PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
C12N 15/10, 15/11, 5/10, 7/04, A61K 39/21, C07B 61/00, G06F 19/00, B01J 19/00 // C12N 15/12, 15/55, 15/49

A2

(11) International Publication Number: **WO 00/18906**

(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22588

(22) International Filing Date: 28 September 1999 (28.09.99)

(30) Priority Data:

60/102,362	29 September 1998 (29.09.98)	US
60/117,729	29 January 1999 (29.01.99)	US
60/118,813	5 February 1999 (05.02.99)	US
60/141,049	24 June 1999 (24.06.99)	US

(71) Applicant *(for all designated States except US)*: MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: PATTEN, Phillip, A. [US/US]; Apartment 506, 2680 Fayette Drive, Mountain View, CA 94040 (US). LIU, Lu [US/US]; 519 Skiff Circle, Redwood City, CA 94056 (US) STEMMER, Willem, P., C. [NL/US]; 108 Kathy Court, Los Gatos, CA 95030 (US).

(74) Agents: QUINE, Jonathan, Alan; Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: SHUFFLING OF CODON ALTERED GENES

(57) **Abstract**

Methods of recombining codon–altered libraries of nucleic acids are provided. The nucleic acids can include conservative or non–conservative modifications of coding sequences, in addition to codon alterations, as compared with wild–type sequences. In addition to making new proteins, methods of generation vectors with reduced rates of reversion to wild–type and attenuated viruses are also provided.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

CROSS REFERENCE TO RELATED APPLICATIONS

This application is a non-provisional filing of USSN 60/102,362, "SHUFFLING OF CODON ALTERED GENES," Attorney Docket No. 02-028500, by Patten and Stemmer, filed 09/29/98, and 60/117,729, "SHUFFLING OF CODON ALTERED GENES," Attorney Docket No. 02-028510, by Patten and Stemmer, filed January 29, 1999. The application is also related to USSN 60/118,813 "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296, filed February 5, 1999; and USSN 60/141,049 "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296-1, filed June 24, 1999.

BACKGROUND

The genetic code is highly degenerate. Every DNA/RNA triplet (codon) encoding an amino acid can typically be altered, with the exception of ATG/AUG (coding for methionine) and TGG/UGG (coding for Tryptophan), without altering the sequence of the protein encoded by the corresponding nucleic acid sequence. Roughly, on average (the distribution of amino acids varies from protein to protein), each coding triplet can be substituted about 3 different ways, since there are 61 codons encoding 20 amino acids (there are 3 additional triplets encoding stop codons, for a total of 64 codons encoding 20 amino acids). This represents a possible sequence diversity of approximately 3^n possible sequences which encode a given protein, where n is the length of the protein in amino acids. As can easily be seen, for proteins of even modest length, the number of possible nucleic acids which can encode the protein exceeds the number of physical particles in the universe (estimated at about 10^{80} particles).

This tremendous potential coding sequence space for individual proteins has interesting evolutionary implications. For example, hypermutable viruses such as HIVs and other retroviruses typically stay one step ahead of the host immune system by accumulating non-random mutations based, in part, upon the particular codons used to encode recognition molecules, e.g., in the envelope portion of the virus. The mutations are non-random because viruses are selected for the ability to mutate to forms which are not quickly recognized by the

host immune system. A consequence of this is that viruses are selected to have a non-random set of codons encoding, e.g., envelope proteins, allowing the viruses to shift forms rapidly by making, e.g., specific point mutations to generate specific alterations in protein structure.

Codon use is also non-random within species. By preferentially making a subset of all possible t-RNAs, cells may conserve energy, and can optimize, or even regulate, the efficiency of cellular translation systems. This fact has long been recognized empirically, often allowing investigators initially to determine the reading frame of a given nucleic acid sequence simply by consideration of the codons resulting from different potential reading frames. One consequence of this "species codon bias" is that proteins within a species have a limited set of possible mutations that can arise as a consequence of, e.g., point mutation. This limits the possible evolution rate of proteins.

In addition to the diversity of nucleic acid coding sequences which encode any given protein, it is now clear that protein sequences are, themselves, quite degenerate. Often, many of the amino acid residues constituting a protein may be substituted for structurally similar amino acid units without significantly changing the tertiary structure of the protein. Thus, it may be difficult to determine which residues to modify or to improve desirable properties of a protein.

For proteins which are commercially valuable, it would be desirable to be able to gain access to a mutational spectrum which is different than that of the native protein. The present invention provides this, and many other features, that will be apparent upon complete review of the following.

SUMMARY OF THE INVENTION

The present invention provides methods of accessing a completely different mutational spectrum for a selected protein than is available in the naturally occurring nucleic acid encoding the protein. This increases the type and rate of forced evolution for the selected protein, allowing for rapid improvement of any detectable characteristic of the protein. In the methods, nucleic acids are synthesized with altered codon usage, and/or which encode one or several amino acid residue changes as compared to the selected protein, where the amino acid and codon usage changes can be conservative or non-conservative. The resulting codon/amino acid modified nucleic acid(s) are recombined using DNA shuffling techniques with either the native nucleic acid, or with each other (or both), typically

using recursive shuffling methods. The nucleic acids or the encoded protein are then screened for a desirable property.

Thus, the invention provides methods of making codon altered nucleic acids. In the methods, a first nucleic acid sequence encoding a first polypeptide sequence is selected. A plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide, or a modified form thereof, are then selected (e.g., a library of codon altered nucleic acids can be selected in a biological assay which recognizes library components or activities), and the plurality of codon altered nucleic acid sequences is recombined to produce a target codon altered nucleic acid encoding a second protein. The target codon altered nucleic acid is then screened for a detectable functional or structural property, optionally including comparison to the properties of the first polypeptide. The goal of such screening is to identify a polypeptide that has a structural or functional property equivalent or superior to the first polypeptide. A nucleic acid encoding such a polypeptide can be used in essentially any procedure desired, including introducing the target codon altered nucleic acid into a cell, vector, virus, attenuated virus (e.g., as a component of a vaccine or immunogenic composition), transgenic organism, or the like.

Kits and compositions for practicing the methods are also provided, including one or more of: cell recombination mixtures and substrates (e.g., nucleic acids with altered codon usage), containers, instructional material for practicing the methods, or the like.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a nucleic acid/amino acid sequence of a part of the monkey EPO gene, which is similar to the human EPO gene.

Figure 2 shows an example of a codon altered EPO nucleic acid sequence.

Figure 3 shows an alignment of naturally occurring EPOs.

Figure 4 is a schematic of the human EPO wobble sequence space.

Figure 5 is a schematic of Mammalian EPO Family-Wobble Sequence Space.

Figure 6 is a sequence alignment of G-CSF homologs, with species information.

Figure 7 is a sequence alignment of G-CSF homologs, with differences broken out.

Figure 8 is a sequence alignment showing the hydrophobic core residues of human G-CSF (blacked out).

Figure 9 is a schematic showing the shuffling strategy for G-CSF.

Figure 10 is a list of oligos used to make a codon altered alkaline phosphatase.

Figure 11 is a map of oligos used to make a codon altered alkaline phosphatase.

Figure 12 is a schematic of vaccination with evolution defective viruses.

Figure 13 is a schematic of different mutations that result from different codon types for ser, arg, and leu.

Figure 14 is a schematic of vaccination with evolution defective viruses.

Figure 15 is a schematic of vaccination with evolution defective viruses showing sophisticated versus non-sophisticated "mutant clouds."

Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Fig. 17 is a schematic of protein evolution with expanded mutation spectra.

Fig. 18, panels A-D show codon altered forms of Env.

Fig. 19 is a list of oligos in one application for synthesis of HIV Env.

DEFINITIONS

Unless clearly indicated to the contrary, the following definitions supplement definitions of terms known in the art.

As used herein, a "recombinant" nucleic acid is a nucleic acid produced by recombination between two or more nucleic acids, or any nucleic acid made by an *in vitro* or artificial process. The term "recombinant" when used with reference to a cell indicates that the cell comprises (and optionally replicates) a heterologous nucleic acid, or expresses a peptide or protein encoded by a heterologous nucleic acid. Recombinant cells can contain genes that are not found within the native (non-recombinant) wild-type form of the cell. Recombinant cells can also contain genes found in the native form of the cell where the genes are modified and re-introduced into the cell by artificial means. The term also encompasses cells that contain a nucleic acid endogenous to the cell that has been artificially modified without removing the nucleic acid from the cell; such modifications include those obtained by gene replacement, site-specific mutation, chimeraplasty, and related techniques.

A "codon altered" nucleic acid is a first nucleic acid that encodes a first polypeptide similar or identical to a naturally occurring polypeptide encoded by a naturally occurring nucleic acid, where the first nucleic acid utilizes a plurality of codons to encode the first polypeptide, which differ from the codons of the naturally occurring nucleic acid that encode the naturally occurring polypeptide.

A "nucleic acid sequence" refers to either a nucleic acid (e.g., RNA, DNA or modified form thereof, in isolated, recombinant or native form) or to a representation of the nucleic acid such as a sequence of letters indicating the primary structure (sequence) of the nucleic acid.

A "polypeptide sequence" refers to either a polypeptide (or modified form thereof, in isolated, recombinant or native form) or to a representation of the polypeptide such as a sequence of letters or other character string information indicating the primary structure (amino acid sequence) of the polypeptide.

A "modified form" of a reference polypeptide is a target polypeptide which has a similar, but not identical, sequence to the reference polypeptide. The sequence of the target polypeptide can differ from the reference polypeptide by conservative or non-conservative substitutions of the reference polypeptide sequence. As noted in more detail, *supra*, different nucleic acids encoding different target polypeptides having different non-conservative substitutions relative to the reference polypeptide can be recombined to produce a recombined nucleic acid encoding a target polypeptide more similar to the reference polypeptide.

A "plurality of forms" of a selected nucleic acid refers to a plurality of homologs of the nucleic acid. The homologs can be from naturally occurring homologs (e.g., two or more homologous genes, or derivatives thereof) or by artificial synthesis of one or more nucleic acids having related sequences, or by modification of one or more nucleic acid to produce related nucleic acids. Nucleic acids are homologous when they are derived, naturally or artificially, from a common ancestor sequence. During natural evolution, this occurs when two or more descendent sequences diverge from a parent sequence over time, i.e., due to mutation and natural selection. Under artificial conditions, divergence occurs, e.g., in one of two ways. First, a given sequence can be artificially recombined with another sequence, as occurs, e.g., during typical cloning, to produce a descendent nucleic acid.

Alternatively, a nucleic acid can be synthesized *de novo*, by synthesizing a nucleic acid which varies in sequence from a given parental nucleic acid sequence.

When there is no explicit knowledge about the ancestry of two nucleic acids, homology is typically inferred by sequence comparison between two sequences. Where two nucleic acid sequences show sequence similarity it is inferred that the two nucleic acids share a common ancestor. The precise level of sequence similarity required to establish homology varies in the art depending on a variety of factors. For purposes of this disclosure, two sequences are considered homologous where they share sufficient sequence identity to allow recombination to occur between two nucleic acid molecules, or when codon changes can be made which would result in two or more nucleic acids having the ability to recombine. Typically, nucleic acids require regions of close similarity spaced roughly the same distance apart to permit recombination to occur.

The terms "identical" or percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or polypeptides refers to two or more sequences or subsequences that have at least about 40%, 50%, 60%, or preferably about 70% or 80% or more, or most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Such "substantially identical" sequences are typically considered to be homologous. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer,

subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al., infra*).

One example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al., supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation

(E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

Another indication that two nucleic acid sequences are substantially identical/ homologous is that the two molecules hybridize to each other under stringent conditions. The phrase "hybridizing specifically to," refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions, including when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences and sequences with higher G:C content remain hybridized at higher temperatures (or at lower salt). An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York.

Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but not to unrelated (non-homologous) sequences.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42 °C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15M NaCl at 72 °C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65 °C for 15 minutes (*see,* Sambrook, *infra.*, for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 1x SSC at 45 °C for 15 minutes. An example low stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 4-6x SSC at 40 °C for 15 minutes. For short probes (*e.g.*, about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 40 °C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. If the signal to noise ratio is less than 2x binding of an unrelated probe (e.g., a nucleic acid encoding a non-homologous protein), the nucleic acids at issue do not hybridize under stringent conditions. Similarly, if the signal to noise ratio is less than 25% as high as that observed for a perfectly matched probe under stringent conditions, the nucleic acids do not "hybridize under stringent conditions" as that term is used herein. This does not apply to highly stringent conditions, as the stringency can theoretically be increased until only a perfectly matched probe will hybridize.

In one example hybridization procedure, a target nucleic acid to be probed is blotted onto a filter by any conventional method. An unrelated nucleic acid such as a plasmid

vector (assuming that the target nucleic acid has no homology with the target nucleic acid) is also blotted, in approximately equal amounts onto the filter. The filter is probed with a labeled probe complementary to the target nucleic acid. The experiment is repeated at gradually increasing stringency of hybridization and wash conditions until signal from the hybridization of the labeled probe to the complementary target is 10-100X as high as to the unrelated plasmid vector nucleic acid. Once these conditions are determined as described above, a test nucleic acid is probed under the same conditions as the target. If signal from the labeled probe is 25% as high or higher than the signal from binding of the probe to the target, the test nucleic acid "hybridizes under stringent conditions" to the probe. If the signal is less than 25% as high, the test nucleic acid does not hybridize under stringent conditions to the probe.

Nucleic acids which do not hybridize to each other under stringent conditions are still recognizable as variant forms of a nucleic acid when the polypeptides they encode are substantially identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. Such nucleic acids are not functionally equivalent, as described in detail herein, due to differences in mRNA folding, alterations of regulatory sequences and the like.

Another indication that two nucleic acid sequences or polypeptides are variant forms is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid, as tested by polyclonal antisera generated to the first polypeptide. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

"Conservatively modified variations" of a particular polynucleotide sequence are those polynucleotide variations that encode identical or essentially identical amino acid sequences, or where the polynucleotide does not encode an amino acid sequence, which encode essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such

nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every polynucleotide sequence described herein which encodes a polypeptide also optionally describes every possible silent variation, except where otherwise noted. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the codon for methionine, and TGG, which is ordinarily the codon for tryptophan) can be modified to yield a peptide which is structurally identical.

Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following five groups each contain amino acids that are conservative substitutions for one another:

Aliphatic: Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I); Aromatic: Phenylalanine (F), Tyrosine (Y), Tryptophan (W); Sulfur-containing: Methionine (M), Cysteine (C);Basic: Arginine (R), Lysine (K), Histidine (H); Acidic: Aspartic acid (D), Glutamic acid (E), Asparagine (N), Glutamine (Q). *See also*, Creighton (1984) *Proteins*, W.H. Freeman and Company. In addition, individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservatively modified variations." Sequences that differ by conservative variations are generally homologous.

The term "isolated", when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular or other components (e.g., library components) with which it is associated in the natural state.

The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.* degenerate codon substitutions) and complementary sequences and as well as the sequence

explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.* (1991) *Nucleic Acid Res. 19*: 5081; Ohtsuka *et al.* (1985) *J. Biol. Chem. 260:* 2605-2608; Cassol *et al.* (1992) ; Rossolini *et al.* (1994) *Mol. Cell. Probes 8*: 91-98). The term nucleic acid is generic to the terms "gene", "DNA," "cDNA", "oligonucleotide," "RNA," "mRNA," and the like.

"Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the gene and detection of such derived products is indicative of the presence and/or abundance of the original gene and/or gene transcript in a sample.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence.

A "recombinant expression cassette" or simply an "expression cassette" is a nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gene in hosts compatible with such sequences. Expression cassettes include at least promoters and optionally, transcription termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (*e.g.*, a nucleic acid encoding a desired polypeptide), and a promoter. Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell. Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

DETAILED DISCUSSION OF THE INVENTION

In the present invention, the sequence diversity of substrates for DNA shuffling procedures is increased by using codon-altered nucleic acids as templates and/or by using templates that encode proteins with conservative or non-conservative amino acid modifications as compared to a selected wild-type protein.

These codon altered nucleic acids can be chemically synthesized (e.g., using standard artificial synthetic protocols, e.g., those typically used by commercial sources from which nucleic acids can be ordered), or can be made using any of a variety of methods herein of available to one of skill. For example, oligonucleotide fragments can be made which correspond to a codon altered nucleic acid which is desired using standard synthetic methods, followed by polymerase and/or ligase mediated oligonucleotide ligation/recombination protocols to generate full-length nucleic acids.

The combination of codon usage modifications and coding modifications can be extensive enough to reduce or, under stringent conditions, even eliminate the hybridization of the codon-altered nucleic acids to a nucleic acid which naturally encodes the selected protein. This dramatically alters the mutations which result from possible single nucleotide mutations, providing access to greater diversity for DNA shuffling protocols.

In addition, the recombination and selection of such nucleic acids during DNA shuffling procedures can result not only in access to a different set of possible mutations, but can also result in modified forms of transcriptional or translational regulation, alterations in nucleic acid localization, mRNA stability and the like. Furthermore, the modified hybridization properties of codon altered nucleic acids leads to alterations in the ability of the nucleic acids to hybridize with potential recombination partners, altering, and ultimately increasing, the available recombination diversity during shuffling.

Furthermore, "family shuffling" using codon-altered substrates even further increases the possible sequence diversity of the starting materials for recombination. As currently practiced, family shuffling methods involve shuffling nucleic acids encoding sequence variants of a given protein (e.g., species or allele homologs). In the present methods, this procedure is modified by generating codon-altered versions of the sequence variants to access additional molecular diversity during recombination. Additional diversity is achieved by conservatively and non-conservatively modifying the starting nucleic acids to encode non-naturally occurring sequence variants. Family shuffling can be performed even using homologs of relatively low identity. In such cases, codons may be changed in one or more of the family members to increase the level of identity between the members, thereby increasing their ability to recombine using the methods of this invention.

Gene shuffling and family shuffling provide two of the most powerful methods available for improving and "migrating" (gradually changing the type of reaction, substrate or activity of a selected protein such as an enzyme, or regulation or structure of an expressed component) the functions of proteins. In family shuffling, homologous sequences, e.g., from different species, chromosomal positions, or due to synthetic alteration, are recombined. In gene shuffling, a single sequence is mutated or otherwise altered and then recombined.

The generation and screening of high quality shuffled libraries provides for DNA shuffling (or "directed evolution"). The availability of appropriate high-throughput analytical chemistry to screen the libraries permits integrated high-throughput shuffling and screening of the libraries to achieve a desired activity.

In one significant embodiment, oligonucleotides for constructing codon-modified nucleic acids are designed in a computer ("in silico"). Predicted codon-modified recombinant nucleic acids can also be determined in silico, i.e., essentially as taught in Selifonov and Stemmer "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

Furthermore, rather than generating codon-modified nucleic acids as substrates for recombination, families of nucleic acids can be recombined simply by appropriate selection of the relevant oligonucleotides which are used in gene reconstruction methods to produce recombinant nucleic acids, i.e., by using codon-modified nucleic acid oligonucleotides as discussed herein in conjunction with family oligonucleotide-mediated shuffling methods, e.g., as taught in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. The technique can be used to recombine homologous or even non-homologous nucleic acid sequences; in the context of the present invention, oligonucleotides corresponding to families of codon-modified nucleic acids are shuffled.

The present invention provides significant advantages over previously used methods for optimization of genes. For example, DNA shuffling of codon modified nucleic acids can result in optimization of a desirable property even in the absence of a detailed

understanding of the mechanism by which the particular property is mediated. In addition, entirely new properties can be obtained upon shuffling of codon modified DNAs, i.e., shuffled DNAs can encode polypeptides or RNAs with properties entirely absent in the parental DNAs which are shuffled. Thus, by modifying the codon usage and/or encoded amino acids of the relevant gene or other nucleic acid, molecular diversity is accessed and sequences can be shuffled to obtain desired, including entirely new, properties.

In general, sequence recombination can be achieved in many different formats and permutations of formats, as described in further detail below.

The targets for modification vary in different applications, as does the property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have enzymatic or therapeutic or other commercially useful activities. A more extensive listing is found *supra*; however, even this list is not intended to be limiting, as essentially any nucleic acid can be codon modified and shuffled, using one or more of the processes herein.

Shuffling methods use at least two variant forms of a starting target (the variant forms can be nucleic acids, or representations thereof, e.g., as character strings in a computer program). The variant forms of candidate codon-altered substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one and preferably at least two positions. The initial diversity between forms can be the result of natural variation, *e.g.*, the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (*e.g.*, allelic variations), or constitute homologs from different organisms (interspecific variants), or constitute artificial homologs, *e.g.*, codon-altered nucleic acids encoding the same or a similar protein. Any or all of these sequences can represent or include codon altered nucleic acids.

Initial diversity can also be induced, *e.g.*, the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (*see,* Liao (1990) *Gene* 88:107-111), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further detail below, and are generally well known). The initial diversity between substrates is greatly augmented in subsequent steps of recombination for library generation.

A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism. The properties or characteristics that can be acquired or improved vary widely, and, of course depend on the choice of substrate.

At least two variant forms of a nucleic acid, e.g., which can confer a desired activity or which can be recombined to produce a desired activity, are recombined to produce a library of recombinant nucleic acids. The library is then screened to identify at least one recombinant nucleic acid that is optimized for the particular property or properties of interest.

Often, improvements are achieved after one round of recombination and selection. However, recursive sequence recombination can be employed to achieve still further improvements in a desired property, or to bring about new (or "distinct") properties. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity. That is, one creates a family of nucleic acid molecules showing some sequence identity to each other but differing due to the presence of mutations. In any given cycle, recombination can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying known methods of mutagenesis (*e.g.*, error-prone PCR or cassette mutagenesis) to either the substrates or products for recombination. In general, however, a single cycle of DNA shuffling of codon-altered nucleic acids provides for generation of surprisingly effective nucleic acids. Accordingly, while recursive approaches to shuffling can be used, single cycle recombination is also preferred. Typically, 2, 3, 4, 5, or even 10 or more cycles of recombination can be performed, each cycle optionally comprising one or more selection steps.

A recombination cycle is usually followed by at least one cycle of screening or selection for molecules having a desired property or characteristic. If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant segments, are sometimes introduced into cells before the screening step. Recombinant segments can also be linked to an appropriate vector or other regulatory sequences before screening.

Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (e.g., bacteriophage) before screening. If recombination is performed *in vivo*, recombination products can sometimes be screened in the cells in which recombination occurred. In other applications, recombinant segments are extracted from the cells, and optionally packaged as viruses, before screening.

The nature of screening or selection depends on what property or characteristic is to be acquired or the property or characteristic for which improvement is sought, and many examples are discussed below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have acquired new or improved properties or characteristics relative to the starting substrates. For example, a gene can have many component sequences, each having a different intended role (*e.g.*, coding sequences, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences can be varied and recombined simultaneously. Screening/selection can then be performed, for example, for recombinant segments that have increased ability to confer activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular screening protocol used for a desired property, initial round(s) of screening can sometimes be performed using bacterial cells due to high transfection efficiencies and ease of culture. However, bacterial expression is often not practical or desired, and yeast, fungal or other eukaryotic systems are also used for library expression and screening. Similarly, other types of screening which are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be performed in the precise cell type of intended use.

If further improvement in a property is desired, at least one and usually a collection of recombinant segments surviving a first round of screening/selection are subject to a further round of recombination. These recombinant segments can be recombined with each other or with exogenous segments representing the original substrates or further variants thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant segments as components of cells, the components can be

subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous screening step identifies desired recombinant segments in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates further recombinant segments which encompass additional diversity that is present in recombinant segments resulting from a previous round (or from multiple previous rounds, e.g., where the process is iteratively repeated).

The second round of recombination can be followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection can be increased between rounds. Also, the nature of the screen and the property being screened for can vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired. Additional rounds of recombination and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

The practice of this invention involves the construction of recombinant nucleic acids and the expression of genes in transfected host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the construction of recombinant nucleic acids such as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, including mutagenesis, include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel")). Methods of transducing cells, including plant and animal cells, with nucleic acids are generally available, as are methods of expressing proteins encoded by such nucleic acids. In addition to Berger, Ausubel and Sambrook, useful general references for culture of animal cells include Freshney (Culture of

Animal Cells, a Manual of Basic Technique, third edition Wiley- Liss, New York (1994)) and the references cited therein, Humason (Animal Tissue Techniques, fourth edition W.H. Freeman and Company (1979)) and Ricciardelli, et al., In Vitro Cell Dev. Biol. 25:1016-1024 (1989). References for plant cell cloning, culture and regeneration include Payne et al. (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York, NY (Payne); and Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) (Gamborg). A variety of Cell culture media are described in Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Additional information for plant cell culture is found in available commercial literature such as the Life Science Research Cell Culture Catalogue (1998) from Sigma- Aldrich, Inc (St Louis, MO) (Sigma-LSRCCC) and, e.g., the Plant Culture Catalogue and supplement (1997) also from Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

Examples of techniques sufficient to direct persons of skill through *in vitro* amplification methods, including the polymerase chain reaction (PCR), the ligase chain reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques (*e.g.*, NASBA) are found in Berger, Sambrook, and Ausubel, *id.*, as well as in Mullis *et al.*, (1987) U.S. Patent No. 4,683,202; PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1, 1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; Kwoh *et al.* (1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli *et al.* (1990) Proc. Natl. Acad. Sci. USA 87, 1874; Lomell *et al.* (1989) J. Clin. Chem 35, 1826; Landegren *et al.*, (1988) Science 241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene 4, 560; Barringer *et al.* (1990) Gene 89, 117, and Sooknanan and Malek (1995) Biotechnology 13: 563-564. Improved methods of cloning *in vitro* amplified nucleic acids are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng *et al.* (1994) Nature 369: 684-685 and the references therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. *See*, Ausbel, Sambrook and Berger, *all supra*.

Oligonucleotides *e.g.*, for use in *in vitro* amplification/ gene reconstruction methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method, e.g., as described by Beaucage and Caruthers (1981), *Tetrahedron Letts.*, 22(20):1859-1862, *e.g.*, using an automated synthesizer, e.g., as described in Needham-VanDevanter *et al.* (1984) *Nucleic Acids Res.*, 12:6159-6168 or as is now practiced routinely in the art. Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill. Purification of oligonucleotides (e.g., using gel-purification methods) to improve the quality of synthesized oligonucleotides can be particularly desirable in the processes herein to improve the quality of nucleic acid synthesis protocols.

As noted, essentially any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (http://www.genco.com), ExpressGen Inc. (www.expressgen.com), Operon Technoloigies Inc. (Alameda, CA) and many others. Similarly, peptides and antibodies can be custom ordered from any of a variety of sources, such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (http://www.htibio.com), BMA Biomedicals Ltd (U.K.), Bio Synthesis, Inc., and many others.

CODON AND AMINO ACID ALTERED LIBRARIES

In the methods of the invention, libraries of codon altered nucleic acids can be made and recombined. The codon altered nucleic acids can also include differences in encoded amino acid sequences, which can be either conservative or non-conservative in nature. The codon altered nucleic acids can be derived from a single parental amino acid sequence, or can be derived from a family of original sequences, *e.g.*, natural or synthetic homologous variants of a given sequence. Libraries can exist, *e.g.*, in pools or aliquots of cells, viral plaques, enzymatically synthesized pools or aliquots of nucleic acids, or chemically synthesisized pools of nucleic acids. Methods of making libraries of nucleic acids are available and taught, *e.g.*, in Berger, Sambrook and Ausubel, *supra*. In one embodiment, a library as used in the invention comprises at least 2 nucleic acid sequences. In additional

embodiments, the libraries of this invention comprise at least 2, 5, 10, 100, 1000, or more nucleic acid sequences.

As applied to the invention, libraries are typically constructed with a high percentage of codons altered relative to an initial (*e.g.*, wild type) nucleic acid. Codon usage divergence for each of the codon altered nucleic acids can be 50%, 75%, or even 90% or more as compared to the first nucleic acid. This eliminates hybridization to the parental nucleic acid (and thereby inhibits recombination with the parental nucleic acid, a desirable feature in certain embodiments discussed below).

In several embodiments of this invention, codons are modified in members of a gene family so as to increase the degree of identity between the members. In one such embodiment, the genes are homologous genes from different species. In such cases, the degree of nucleic acid identity may be lower than the degree of amino acid identity, at least in part, because of differences in codon usage between the species. In additional embodiments, the homologous genes represent different members of a gene family within a single species. Such genes may encode functionally distinct members of a gene family that nevertheless share significant structural or functional similarity. In preferred embodiments, homologous genes are reverse translated into nucleic acid sequences, and the nucleic acid sequences are modified so as to increase the level of identity between them. Nucleic acids with the modified sequences can then be synthesized *in vitro*. In particularly preferred embodiments, the modified nucleic acid sequences are at least as identical to each other as the original amino acid sequences.

Additional sequence diversity is provided by generating nucleic acids with non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid. This provides for reversion to wild-type upon recombination, while optionally allowing for the incorporation of non-conservative changes to the sequence in the event that they produce a detectable improvement during screening.

Modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide is also provided. This modification of core amino acids provides minor differences in encoded proteins, while changing the mutational spectrum of the resulting nucleic acid, thereby increasing sequence diversity.

In addition, due to the constraints of the translational machinery for a given cell, codon usage may need to be altered when expressed sequences are shuttled between different organisms (e.g., animal cells, plant cells, bacterial cells, etc.) for optimal expression. This produces a nucleic acid which encodes the same protein, but which, after typical forms of point mutation, will access a different mutational diversity than the original form of the protein.

In one embodiment, phage libraries are made and recombined in mutator strains such as cells with mutant or impaired gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. High multiplicity of infection (MOI) libraries are used to infect the cells to increase recombination frequency. Additional strategies for making phage libraries and or for recombining DNA from donor and recipient cells are set forth in U.S. Pat. No. 5,521,077. Additional recombination strategies for recombining plasmids in yeast are set forth in WO 97 07205.

The library to be made can be an *in vitro* set of molecules, or present in cells, phage or the like. Virtual libraries of nucleic acids generated in silico are also a feature of the invention (*see also*, Selifonov and Stemmer, *supra*). Generally, the library is screened to identify at least one recombinant nucleic acid that exhibits distinct or improved activity compared to the parental nucleic acid or nucleic acids which are recombined. Additional details on making appropriate libraries are found below, *e.g.*, in the section entitled "Formats for Sequence Recombination."

TARGETS FOR CODON MODIFICATION AND SHUFFLING

Essentially any nucleic acid can be codon altered and shuffled. No attempt is made herein to identify the hundreds of thousands of known nucleic acids. Common sequence repositories for known proteins include GenBank EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

One class of preferred targets for activation includes nucleic acids encoding therapeutic proteins such as erythropoietin (EPO), insulin, peptide hormones such as human growth hormone; growth factors and cytokines such as epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1β, MCP-1, epidermal growth

factor, fibroblast growth factor, hepatocyte growth factor, insulin-like growth factor, the interferons, the interleukins, keratinocyte growth factor, leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF etc. Many of these proteins are commercially available (*See,* e.g., the Sigma BioSciences 1997 catalogue and price list), and the corresponding genes are well-known.

Another class of preferred targets are transcriptional and expression activators. Example transcriptional and expression activators include genes and proteins that modulate cell growth, differentiation, regulation, or the like. Expression and transcriptional activators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, and animals, including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, e.g., by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and degrading RNA. Expression activators include cytokines, inflammatory molecules, growth factors, their receptors, and oncogene products, e.g., interleukins (e.g., IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, e.g., Mos, Ras, Raf, and Met; and transcriptional activators and suppressors, e.g., p53, Tat, Fos, Myc, Jun, Myb, Rel, and steroid hormone receptors such as those for estrogen, progesterone, testosterone, aldosterone, the LDL receptor ligand and corticosterone.

Similarly, proteins from infectious organisms for possible vaccine applications, described in more detail below, including infectious fungi, e.g., *Aspergillus*, *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (e.g., *aureus*), *Streptococci* (e.g., *pneumoniae*), *Clostridia* (e.g., *perfringens*), *Neisseria* (e.g., *gonorrhoea*), *Enterobacteriaceae* (e.g., *coli*), *Helicobacter* (e.g., *pylori*), *Vibrio* (e.g., *cholerae*), *Capylobacter* (e.g., *jejuni*), *Pseudomonas* (e.g., *aeruginosa*), *Haemophilus* (e.g., *influenzae*), *Bordetella* (e.g., *pertussis*), *Mycoplasma* (e.g., *pneumoniae*), *Ureaplasma* (e.g., *urealyticum*), *Legionella* (e.g., *pneumophila*), *Spirochetes* (e.g., *Treponema*, *Leptospira,* and *Borrelia*),

Mycobacteria (e.g., *tuberculosis*, *smegmatis*), *Actinomyces* (e.g., *israelii*), *Nocardia* (e.g., *asteroides*), *Chlamydia* (e.g., *trachomatis*), *Rickettsia*, *Coxiella*, *Ehrilichia*, *Rochalimaea*, *Brucella*, *Yersinia*, *Fracisella*, and *Pasteurella*; protozoa such as sporozoa (e.g., *Plasmodia*), *rhizopods* (e.g., *Entamoeba*) and flagellates (*Trypanosoma*, *Leishmania*, *Trichomonas*, *Giardia*, etc.); viruses such as (+) RNA viruses (examples include Poxviruses e.g., *vaccinia*; Picornaviruses, e.g. *polio*; Togaviruses, e.g., *rubella*; Flaviviruses, e.g., HCV; and Coronaviruses), (-) RNA viruses (examples include Rhabdoviruses, e.g., VSV; Paramyxovimses, e.g., RSV; Orthomyxovimses, e.g., influenza; Bunyaviruses; and Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, i.e., Retroviruses, e.g., especially HIV and HTLV, and certain DNA to RNA viruses such as Hepatitis B virus.

Other proteins relevant to non-medical uses, such as inhibitors of transcription or toxins of crop pests e.g., insects, fungi, weed plants, and the like, are also preferred targets for shuffling. Industrially important enzymes such as monooxygenases, proteases, nucleases, and lipases are also preferred targets. As an example, subtilisin can be evolved by shuffling codon altered forms of the gene for subtilisin (von der Osten et al., J. Biotechnol. 28:55-68 (1993) provide a subtilisin coding nucleic acid). Proteins which aid in folding such as the chaperonins are also preferred.

Preferred known genes suitable for codon alteration and shuffling also include the following: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, Apolipoprotein, Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Atrial peptides, C-X-C chemokines (e.g., T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4, SDF-1, PF4, MIG), Calcitonin, CC chemokines (e.g., Monocyte chemoattractant protein-1, Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309, R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, Collagen, Colony stimulating factor (CSF), Complement factor 5a, Complement inhibitor, Complement receptor 1, Factor IX, Factor VII, Factor VIII, Factor X, Fibrinogen, Fibronectin, Glucocerebrosidase, Gonadotropin, Hedgehog proteins (e.g., Sonic, Indian, Desert), Hemoglobin (for blood substitute; for radiosensitization), Hirudin, Human serum albumin, Lactoferrin, Luciferase, Neurturin, Neutrophil inhibitory factor (NIF), Osteogenic protein,

Parathyroid hormone, Protein A, Protein G, Relaxin, Renin, Salmon calcitonin, Salmon growth hormone, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, i.e., Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Toxic shock syndrome toxin (TSST-1), Exfoliating toxins A and B, Pyrogenic exotoxins A, B, and C, and M. arthritides mitogen, Superoxide dismutase, Thymosin alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha) and Urokinase. Many other known coding nucleic acids, such as those in Genebank™, can be codon-altered and shuffled.

GENES WITH CODON USAGE REDESIGNED AND CHEMICALLY SYNTHESIZED AS STARTING MATERIALS FOR GENE FAMILY SHUFFLING--EXPANDING THE DIVERSITY OF DNA SHUFFLING.

Because the genetic coding preference among organisms ranges from quite similar to very different, homologous genes from different organisms can have significantly lower homology at the nucleic acid level than at the amino acid level. For example, genetic information for some bacterial species is high in GC content (up to 70%), while others have AT rich (>60%) codon usage. Thus, genes from different organisms may have, for example, 40-60% amino acid identity but only 25-35% nucleic acid identity. It is often desirable to increase such levels of nucleic acid identity so as to enhance the ability of the homologous sequences to recombine, thereby increasing the efficiency of family shuffling using the methods of this invention. In other aspects, it is actually preferable to decrease the rate of recombination in a system, e.g., when using vectors it is sometimes desirable to decrease the rate of recombination between the vector and the host DNA, thereby increasing the safety of the vector. The following examples address specific issues with regard to shuffling codon altered nucleic acids.

Altering codon usage to increase homology

In one aspect, protein sequences of gene family members are reverse translated back into DNA sequences, for example by using one of the preferable codon usage charts in any conventional DNA manipulation program (*e.g.* the Wisconsin Package™, SeqWeb, OMIGA, SeqApp, SeqPup, MacVector, DNA stryder, GeneWorks, etc.). The

choice of codon usage is often determined by the host in which the genes will be expressed. After maximizing the percentage of DNA sequence identity, the genes are chemically synthesized, *e.g.*, using a high throughput oligonucleotide synthesizer in, *e.g.* a 96-well format, optionally in conjunction with polymerase and/ or ligase gene synthesis methods.

In general, the DNA sequence similarity after such treatment will be at least as high as the amino acid similarity, but can be at least about 10% to 15% *higher* than the amino acid identity (in contrast to the situation for naturally occurring genes, which are ordinarily less well conserved than encoded polypeptides), based on the random frequency of sequence identity for any given codon. In most cases, the minimal requirement for amino acid identity can be as low as about 35% while still retaining adequate nucleic acid homology for standard recombination methods (as discussed, *supra*, oligonucleotide-mediated recombination methods do not require high levels of similarity to achieve recombination). In some cases, however, the minimal amino acid identity can be even lower, *e.g.* if the conserved regions are clustered within the genes.

Example: Shuffling codon-modified EPO

The protein erythropoietin alpha, also known as EPO, Epogen, and Procrit is a hematopoietic hormone, providing a variety of benefits to patients suffering from anemia (a common symptom of, e.g., AIDS). EPO is produced as a pharmaceutical, with sales of nearly 1 billion dollars world-wide. Accordingly, proteins with EPO-like activity (and preferably superior activity) are of substantial commercial interest.

Figure 1 shows the sequence of a part of the monkey EPO gene, which is similar to the human EPO gene. Figure 2 shows an example of a codon altered EPO nucleic acid (or "wobble" EPO gene). In general, transversions rather than transition mutations are made where possible. The purpose of this strategy is to maximally disrupt hybridization of the resulting gene with naturally occurring EPOs. Figure 3 shows an alignment of naturally occurring EPOs.

This strategy is further fine-tuned by applying standard rules of base pairing (e.g., elimination of G-C pairing and GC stacking) to maximize sequence disruption; in addition, conservative or non-conservative amino acid modifications can also be made (in some cases, where multiple codon-altered nucleic acids are shuffled, it is desirable to make codon altered nucleic acids with non-overlapping non-conservative substitutions to permit

reversion to the wild-type amino acid during shuffling). The size of the sequence space for nucleic acids encoding EPO is large, at about 2.8 x 10^{88} different sequences (there are about 10^{80} particles in the universe; thus, it is physically impossible to make all of the possible sequences encoding EPO). As indicated schematically in Figure 4, if one only considers the maximally divergent wobble genes (those that use the alternative types of codons for leucine, arginine, and serine), there is still a sequence space of 10^{38} sequences encoding EPO. The overall strategy is to synthesize a library of wobble genes, screen for expression and activity and DNA shuffle desirable genes as desired (e.g., by recursive processes).

It is of interest to further evolve codon-altered nucleic acids. Shuffling with other homologous genes from nature, designed genes (incorporating libraries of designed sequence variation), and genes containing mutations of interest are strategies for evolving any gene of interest. However, the codon altered nucleic acid may not be easily shuffled with these genes because of the sequence differences; or they may be undesirable for other reasons (e.g., the naturally occurring sequences may be proprietary, or include proprietary elements). These difficulties can be avoided by synthesizing codon-altered homologous nucleic acids which encode desired amino acid variations (e.g., those found in homologous genes), but which have a codon-set close to the nucleic acid(s) to be recombined (thereby permitting, e.g., hybridization during recombination).

For example, after identifying homologues of interest (e.g., those shown in Figure 3 for EPO), codon-altered nucleic acids encoding the same proteins are synthesized with a similar codon selection. Standard family shuffling is then practiced with the codon altered nucleic acids. This is shown schematically for EPO in Figure 5.

EPO wobble variants are screened for expression and then receptor binding assays are conducted in an ELISA format, using human EPOr-Fc fusions. Following selection of binding variants, activity is measured as thymidine incorporation in UT7-EPO (A human bone marrow cell line) cell proliferation assays. Cells are treated for 2-3 days with various concentration of EPO variants after which time they are incubated in the presence of 3-H thymidine for 4 hours and incorporation of thymidine is measured. *See also*, Erickson-miller et al. (1997) Blood 90:2421 (for the receptor binding assay), and Wen et al. (1994) J. Biol. Chem. 269:22839-22846 (for the thymidine incorporation assay).

Assays for selecting EPO can also be based, e.g., on the ability of EPO proteins to stimulate the growth of blood cell, e.g., in vitro or in vivo.

Example: Codon Shuffling G-CSF

Family shuffling can be used to breed diversity from genes into the libraries to be screened. Additionally, design heuristics such as randomization of hydrophobic core residues can be used to take advantage of the redundancy between primary structure and tertiary structure of proteins (i.e. many different primary structures encode proteins with very similar three dimensional structures).

Design heuristics are employed to create a sequence space of mutants that are predicted to be highly biased (relative to random mutagenesis) to encode proteins which preserve the original activity. Methods such as high throughput (HTP) screening and phage panning are used to identify members of the designed libraries that have the desired activity. DNA shuffling is used to breed this population of active clones in order to fine tune the mutants, thus allowing one to evolve variants with equivalent or superior function relative to the naturally occurring proteins.

Figures 6 and 7 show several mammalian homologues of G-CSF. Figure 8 shows the hydrophobic core residues of human G-CSF (blacked out). Figure 9 shows a strategy for evolving variants of human G-CSF that are highly divergent in sequence. First, three genes are synthesized (Genes 1, 2 and 3, Figure 8) which contain all of the mammalian homologue diversity of G-CSF. These genes are shuffled, phage panned against the G-CSF receptor, and HTP screened for biological function (receptor activation). Active clones are iteratively shuffled and screened if necessary to give evolved variants that rival or surpass the human gene in activity (on human cells).

Next, one evolves a variant that has a highly mutated hydrophobic core. This is schematically illustrated in Figure 8, and the specific strategy for performing the biological screening is schematically illustrated in Figure 9. it is expected that the best mutants obtained after screening hydrophobic core randomized libraries may be less active than wild type human G-CSF because it is difficult to initially optimize activity in such a procedure. Family shuffling is used to obtain optimized variants. This is done by synthesizing genes which contain mammalian homologue diversity at all but the hydrophobic core positions; but

they are synthesized in the context of an evolved, non-wild type hydrophobic core. Family shuffling is used to optimize around the new hydrophobic core.

This strategy works because there are functionally similar hydrophobic cores for wild type proteins that consist of largely different amino acids than the wild type protein. This understanding is supported by recent experiments in model systems. For example, 53% of randomized sequences for three residues in the hydrophobic core of lambda repressor are folded and biologically active (Lim and Sauer (1991) J. Mol. Biol. 219:359-376). Protein design by patterning of polar and non-polar amino acids, where 24 residues in the hydrophobic core of a 4-helix bundle protein were randomized by Kamtekar et al. (1993) Science 262:1321. Folded, alpha helical proteins were recovered from about 1% of the clones. Desjarlais and Handel (1995) Current Opinion in Biotechnology 6:460-466 showed a mutant of Rop, another 4-helix bundle protein, where four hydrophobic core residues have been randomized and active mutants have been obtained. Axe et al. (1996) PNAS 95:5590-5594 showed that randomizing 13 hydrophobic core residues in the enzyme barnase resulted in 23% of the clones in the library retaining biological activity. Gassener et al. (1996) PNAS 93:12155-12158 describe a mutant of T4 lysozyme where 10 residues in the hydrophobic core are replaced with Met. This is taken as evidence that the hydrophobic core of this protein is very tolerant to substitution. Taken together, this experimental evidence on model systems shows that the hydrophobic cores of many proteins can be replaced with other hydrophobic residues that pack in a similar fashion to give an active protein. This degeneracy is exploited to evolve novel forms of natural and codon-altered genes.

A related approach is to search the protein databases for a protein that has a similar activity to a protein that on wishes to evolve. Denesyuk et al. (1996) J. Theor. Biol shows the results of such a search for G-CSF. LIF is a very similarly folded protein. One can use LIF as a `scaffold' on which to place residues of G-CSF that are required for activity. Given LIF with a G-CSF "toupee," one would family shuffle the LIF scaffold so as to obtain a variant in which the toupee is displayed in a fully biologically active form.

Another approach is to use computational methods to create families of variants that are predicted to be functional. Dahiyat and Mayo Science recently described computer methods that are used to design proteins. Proteins are simulated on the computer, often with the aid of genetic algorithms, and a subset that are deemed `fit' are actually

synthesized and 'analyzed'. These computational methods are becoming increasingly powerful. They would be useful to, for example, predict a family of mutations on the surface of a protein that would not destroy function. DNA shuffling can be used to optimized active clones obtained by design. Taking the example of G-CSF, one could use computational methods in combination with all structure function data (for example alanine scan data for G-CSF reported recently by Reidhaar-Olsen in Biochemistry) to design a family of putatively functional variants. One could, for example, design the family to have minimal DNA identity to the wild type gene given the design constraints. This library is synthesized, put through biological screens and/or selections (i.e. panning against the G-CSF receptor), and active variants are obtained. DNA shuffling is then used to evolve these active variants to have the desired level of function.

G-CSF proteins are displayed on phage and screened for binding to human G-CSF receptor in an ELISA format. Variants that bind receptor are selected in a high throughput screen for receptor activation. This cell based assay measures receptor activation via a reporter gene (such as luciferase) activated by a G-CSF responsive construct containing STAT binding elements. Cells (such as HepG2) are transformed with a G-CSF responsive reporter plasmid and treated with the codon shuffled G-CSF variant for 2.5 hours. Cells are then lysed and luciferase activity measured. *See also*, Tian et al (1998) Science 281:257-259.

Example: Codon Shuffling Alkaline Phosphatase

Alkaline phosphatase is a widely used reporter enzyme for ELISA assays, protein fusion assays, and in a secreted form as a reporter gene for mammalian cells. A more active form of the enzyme is desirable.

A codon altered form of alkaline phosphatase was generated by PCR assembly using the oligos set forth in Figure 10. A map of the oligos is set forth in Figure 11. The procedure used was essentially identical to that taught in Stemmer et al. (1994) Gene 164:49-57. In brief, the oligos were mixed 1:1 at a variety of dilutions and PCR assembled by performing e.g., 25-60 cycles of PCR at e.g., 94 °C (60 sec.), 94 °C (30 sec.), 50 °C (30 sec.), 72 °C (30 sec). Assembly of the BIAP gene was conducted in a circular format and gene fragements were purified. ~100,000 colonies were screened on LB/am plates (~1/10 are wt plasmid). About 1/10 showed a bluer color than background. Plasmid DNA showed a correct insertion.

In general, petri-dish screening using the typical colorimetric assay for phosphatase activity can be used for screening. This has the advantage or being simple, high throughput, and semi quantitative. Microtiter plate screening, also preferred, is colorimetric, and quantitative, although additional instrumentation can be required for implementation.

Example: Codon Shuffling to Reduce Competent Virus Production from Vectors and to Generate Attenuated Viruses as Immunogenic Compositions and Vaccines

Cells can be stably transduced with a number of viral vectors including those derived from retroviruses, pox viruses, adenoviruses (Ads), herpes viruses and parvoviruses. Common viral vectors include those derived from murine leukemia viruses (MuLV), gibbon ape leukemia viruses (GaLV), human immuno deficiency viruses (HIV), adenoviruses, adeno associated viruses (AAVs), Epstein Barr viruses, canarypox viruses, cowpox viruses, and vaccinia viruses. Viral vectors based upon retroviruses, adeno-associated viruses, herpes viruses and adenoviruses are all used as gene therapy vectors for the introduction of therapeutic nucleic acids into the cells of an organism by *ex vivo* and *in vivo* methods.

When using viral vectors, packaging cells are commonly used to prepare virions used to transduce target cells. In these vectors, trans-active genes are rendered inactive and "rescued" by trans-complementation to provide a packaged vector. This form of trans complementation is provided by co-infection of a packaging cell with a virus or vector which supplies functions missing from a particular gene therapy vector in trans, or by using a cell line (e.g., 293 cells) which have viral components integrated into the genome of the packaging cell. For instance, cells transduced with HIV or murine retroviral proviral sequences which lack the nucleic acid packaging site produce retroviral trans active components, but do not specifically incorporate the retroviral nucleic acids into the capsids produced, and therefore produce little or no live virus.

If these transduced "packaging" cells are subsequently transduced with a vector nucleic acid which lacks coding sequences for retroviral trans active functions, but includes a packaging signal, the vector nucleic acid is packaged into an infective virion. A number of packaging cell lines useful for MoMLV-based vectors are known in the art, such as PA317 (ATCC CRL 9078) which expresses MoMLV core and envelope proteins see, Miller et al. J. Virol. 65:2220-2224 (1991). Carrol et al. (1994) Journal of virology 68(9):6047-6051 describe the construction of packaging cell lines for HIV viruses.

Reciprocal complementation of defective HIV molecular clones is described, e.g., in Lori et al. (1992) Journal of Virology 66(9) 5553-5560.

Functions of viral replication not supplied by trans-complementation which are necessary for replication of the vector are present in the vector. In HIV, this typically includes, e.g., the TAR sequence, the sequences necessary for HIV packaging, the RRE sequence if the instability elements of the p17 gene of gag is included, and sequences encoding the polypurine tract. HIV sequences that contain these functions include a portion of the 5' long terminal repeat (LTR) and sequences downstream of the 5' LTR responsible for efficient packaging, i.e., through the major splice donor site ("MSD"), and the polypurine tract upstream of the 3' LTR through the U3R section of the 3' LTR. The packaging site (psi site or ψ site) is partially located adjacent to the 5' LTR, primarily between the MSD site and the gag initiator codon (AUG) in the leader sequence. See, Garzino-Demo et al. (1995) Hum. Gene Ther. 6(2): 177-184. For a general description of the structural elements of the HIV genome, see, Holmes et al. PCT/EP92/02787.

Another common vector is based upon adenovirus. Typically, vectors which include the adenovirus ITRs (Gingeras *et al.* (1982) *J. Biol. Chem.* 257:13475-13491) are packaged in, e.g., 293 cells, which provide many of the components necessary for vector packaging.

Adeno-associated viruses (AAVs) utilize helper viruses such as adenovirus or herpes virus to achieve productive infection. In the absence of helper virus functions, AAV integrates (site-specifically) into a host cell's genome, but the integrated AAV genome has no pathogenic effect. The integration step allows the AAV genome to remain genetically intact until the host is exposed to the appropriate environmental conditions (*e.g.*, a lytic helper virus), whereupon it re-enters the lytic life-cycle. Samulski (1993) *Current Opinion in Genetic and Development* 3:74-80 and the references cited therein provides an overview of the AAV life cycle. For a general review of AAVs and of the adenovirus or herpes helper functions *see*, Berns and Bohensky (1987) *Advanced in Virus Research*, Academic Press., 32:243-306. The genome of AAV is described in Laughlin *et al.* (1983) *Gene*, 23:65-73. Expression of AAV is described in Beaton *et al.* (1989) *J. Virol.*, 63:4450-4454. In general, the packaging sites for all parvoviruses, including B 19 and AAV are located in the viral ITRs. Recombinant AAV vectors (rAAV vectors) deliver foreign nucleic acids to a wide

range of mammalian cells (Hermonat & Muzycka (1984) *Proc Natl Acad Sci USA* 81:6466-6470; Tratschin *et al.* (1985) *Mol Cell Biol* 5:3251-3260), integrate into the host chromosome (McLaughlin *et al.* (1988) *J Virol* 62: 1963-1973), and show stable expression of the transgene in cell and animal models (Flotte *et al.* (1993) *Proc Natl Acad Sci USA* 90:10613-10617). rAAV vectors are able to infect non-dividing cells (Podsakoff *et al.* (1994) *J Virol* 68:5656-66; Flotte *et al.* (1994) *Am. J. Respir. Cell Mol. Biol.* 11:517-521). Further advantages of rAAV vectors include the lack of an intrinsic strong promoter, thus avoiding possible activation of downstream cellular sequences, and the vector s naked icosohedral capsid structure, which renders the vectors stable and easy to concentrate by common laboratory techniques.

One problem with previously existing vector packaging strategies is that vectors to be packaged can recombine with nucleic acids providing packaging functions in trans, producing a replication-competent virus. This can be a problem both when vectors are produced for therapeutic applications (e.g., in gene therapy) and during production of encoded components in vitro. The present invention provides a way of reducing or eliminating recombination between nucleic acids encoding trans-active components and vector nucleic acids encoding packaging sites.

In particular, nucleic acid subsequences of a vector which are adjacent to modified or deleted elements provided in trans, are codon modified to eliminate hybridization to wild-type sequences. Because these sequences do not hybridize, they cannot recombine with nucleic acids producing trans-active components. One additional advantage of this approach is that the vectors also cannot recombine with live viruses, e.g., in a human body which is infected with a virus that packages vector elements. As noted above, two types of gene therapy vectors are those based upon retroviruses (which can be packaged by, e.g., HIV-1) and adenoviruses (which can be packaged by adenovirus).

Alternatively, the nucleic acids encoding trans-active components can be codon modified so that they do not hybridize to wild-type sequences. This also prevents recombination with vectors having wild-type sequences, preventing recombination and formation of replication competent viruses.

After codon modification, vectors or trans active nucleic acids can be shuffled as described *supra*, and screened for the ability to package nucleic acids, or to be packaged, as appropriate.

It will be appreciated that codon modification of viral sequences has an additional use as well. Codon alteration of viral sequences can result in attenuation of the virus, e.g., due to modification of regulatory sequences, alterations in mRNA secondary structure, inefficient translation due to rare codon use, and the like. Such "codon attenuated" viruses have a significant advantage over existing attenuated viruses (which are typically generated by serial passage in cells other than the normal host type for the virus). In particular, codon attenuated viruses can encode a wild-type set of proteins, making them ideal as immunogenic compositions to generate antibodies, or to use as vaccines. Viral proteins can also be used in various diagnostic assays. For example, the standard diagnostic test for HIV infection in current use tests for the presence of anti-HIV antibodies in blood by probing with viral proteins.

Example: Codon Usage Libraries to evolve Functional Variants with reduced Recombination With Natural Gene Sequences--Adenovirus

Adenovirus is a common vector used, e.g., for gene therapy. The virus is typically modified to make it replication deficient. This can be achieved e.g., by deleting the E1 and E4 genes. The functions of E1 and E4 can be supplied by trans complementation when E1 and E4 deleted vectors are grown in the ubiquitous human embryonic kidney cell line 293, which has uncharacterized adenovirus fragments incorporated into their genome that supply the missing functions in trans. The replication defective adenoviral vectors recombine at a low, but clinically significant frequency, resulting in replication competent adenovirus contamination of vector preparations. Because adenovirus has detrimental effects on health, this is a significant problem for application of adenovirus-based gene therapy vectors.

In the present invention, a codon usage library encompassing several hundred bases to several kilobases of sequence flanking the adenovirus E1 and E4 genes are made. The library is designed to enforce a high degree of divergence from the natural adenoviral consensus sequence, while at the same time incorporating a large degree of degeneracy in the codons to allow for a large space of sequence diversity to be searched. The design principle

is to obtain mutants that encode the same or similar protein sequence, but with many mismatches to the wild-type E1 and E4 sequences found in the 293 genome. These mismatches strongly reduce the frequency of unwanted recombination with the trans complementary genes. Consequently, engineered adenoviral vectors, or adenovirus helper vectors which package adenoviral sequences which include packaging sequences (adenoviral or adeno-associated viral ITRs) in trans have reduced levels of recombination. This provides for a lower rate of competent adenovirus production, making culture and production of such vectors safer.

Evolution Impaired Viruses Created by Massive Codon Usage Alteration As a General Approach to Vaccines--HIV

HIV-1 and HIV-2 are genetically related, antigenically cross reactive, and share a common cellular receptor (CD4). *See*, Rosenburg and Fauci (1993) in *Fundamental Immunology, Third Edition* Paul (ed) Raven Press, Ltd., New York (Rosenburg and Fauci 1) and the references therein for an overview of HIV infection. HIV-1 infection is epidemic world wide, causing a variety of immune system-failure related phenomena commonly termed acquired immune deficiency syndrome (AIDS). HIV type 2 (HIV-2) has been isolated from both healthy individuals and patients with AIDS-like illnesses (Andreasson, *et al.* (1993) *Aids* 7, 989-93; Clavel, *et al.* (1986) *Nature*, 324, 691-695; Gao, *et al.* (1992) *Nature* 358, 495-9; Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Kanki, *et al.* (1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30; Naucler, *et al.* (1993) *International Journal of STD and Aids* 4, 217-21; Naucler, *et al.* (1991) *Aids* 5, 301-4). Although HIV-2 AIDS cases have been identified principally from West Africa, sporadic HIV-2 related AIDS cases have also been reported in the United States (O'Brien, *et al.* (1991) *Aids* 5, 85-8) and elsewhere. HIV-2 will likely become endemic in other regions over time, following routes of transmission similar to HIV-1 (Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30). Epidemiological studies suggest that HIV-2 produces human disease with lesser penetrance than HIV-1, and exhibits a considerably longer period of clinical latency (at least

25 years, and possibly longer, as opposed to less than a decade for HIV-1; *see*, Kanki, *et al.* (1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30, and Travers *et al.* (1995) *Science* 268: 1612-1615).

The ability of HIV virus populations to rapidly point mutate to avoid the immune response poses a special challenge for vaccine design. While the immune system has responded to viruses in a gradual and co-evolutionary manner, the present invention provides a general approach that provides for massively faster evolution to produce new vaccines to stimulate more effective immune responses.

For example, during the incubation period for HIV infection, which lasts for several years, low titers of HIV can result from high HIV replication rates in conjunction with efficient viral clearance by the immune system. In response to these selective forces, virus mutations are selected which reduce recognition and neutralization by the immune system's B and T-cell responses. *See,* Lukashov et al. (1995) J. Virol. 69:6911-6916. During the long incubation time, these mutations accumulate and eventually overwhelm the immune system's defenses. *See,* Ho et al. (1995) Nature.

Live attenuated vaccines, typically produced by prolonged growth of human viruses in animal cells, have proven useful as vaccines for several diseases, including mumps, rubella and measles. Attenuation involves the slow accumulation of many mutations throughout the viral genome during the course of adapting to growth in the animal cells. When used to vaccinate humans, the attenuated virus grows only weakly and elicits a complex immune response which the virus is unable to avoid. The mutations in the attenuated virus could, in principle, revert in the same stepwise fashion that it underwent to grow in culture.

The risk of reversion is highest in viruses with a high mutation rate such as HIV-1, which makes this strategy dangerous under current techniques for vaccine development. It is worth noting, however, that protective effects against HIV-1 are observed following infection with the related HIV-2 virus, which is much less pathogenic than HIV-1. Thus, protective effects against HIV can be achieved with live vaccines.

To reduce the risk of reversion, a large number of mutations need to accumulate in the virus. However, if too many mutations are present, the immune system in

effect recognizes the attenuated virus, but not the virus against which a protective effect is sought.

As provided herein, immunogenic compositions such as vaccines are created which contain a large number of silent substitutions. In contrast to existing attenuated viruses, such viruses have native protein sequences and elicit essentially the same immune responses as the corresponding wild-type virus (typically one or a few additional disabling mutations can also be incorporated). Codon alteration results in two effects that both increase the potential of the vaccine.

First, like standard attenuated viruses, the growth of codon-altered viruses is attenuated, due to the effect of the codon alterations on translation, regulatory sequences, mRNA folding, packaging, and the like. For example, regulation of HIV-1 envelope expression has been observed as a result of codon usage. *See,* Haas et al. (1989) Current Biology 6(3):315-324.

Second, codon alteration results in impairment of virus evolution. As discussed above, modification of the codons alters the mutational escape spectrum of the virus, upsetting the evolutionary selection for specific codons.

The six codon amino acids are the best targets for codon alteration. Serine, arginine and leucine each have one group of four codons, plus two codons in an unrelated group. *See*, Figure 12. Switching all of the serine codons from AGY to TCX and vice versa, yields proteins with unaltered amino acid sequences. *See also*, Figure 13. However, these codon groups differ significantly in the spectrum of the amino acids that they yield upon point mutation. Of all possible point mutations of one codon for serine (TCA) 78% result in a different amino acid compared to point mutations obtained for the AGT codon for serine. *See*, Figure 13. A virus with hundreds of codon alterations is in, statistically, a very different mutational space, able to access a totally different mutation spectrum, or "cloud," compared to the wild-type virus. The overall strategy for producing an evolution-defective virus is additionally set forth in Figures 14, 15 and 17. Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Point mutation is critical for viruses such as HIV-1 to stay ahead of the host immune system. The amino acid mutations that are required for virus escape are likely not random. Wild type codon usage has evolved to allow optimal immune system evasion. The

wild type codon usage is likely to favor mutations that represent alterations that avoid the host immune system, without detrimentally affecting the protein(s) encoded. While complex, this natural pattern of amino acid sequence change of the natural virus in response to the host system is non-random and weakly predictable. *See also*, Seiller-Moiseiwitsch et al. (1994) Annu. Rev. Genet. 28:559-596.

Changing all of the codons for ser, arg and leu in the 875 aa envelope polyprotein of HIV-1 (e.g., strain MN) would affect 187 codons (22%) resulting in 561 mutations. *See*, e.g., Figure 18, panels A-D. If all of the HIV proteins were altered, the number of mutations would be more than three-fold higher. The construction of such codon modified viruses is simplified by recent advances in the synthesis of long DNA sequences, which enable the assembly of a plasmid of average size from 40 mer oligos in a single step with about 75% efficiency. *See*, Stemmer (1994) Nature 370_389-391. *See also*, Figure 19 for a list of oligos in one application for synthesis of HIV Env. While the synthesis of the envelope gene is sufficient, synthesis of the whole HIV genome from oligos can be performed by this method.

In practice, a preferred balance of attenuation and evolution impairment is obtained by DNA shuffling (e.g., Stemmer et al. (1995) Gene 164:49-53), e.g., of the wild-type and codon altered sequences, followed by selection of the resulting library of viruses that retain moderate growth despite many codon alterations.

While attenuation that can be obtained by this approach may be sufficient for obtaining a vaccine for most viruses, for HIV-1, the evolution impairment is more important, due to the high mutation rate of the virus. Live vaccines are used only if they elicit an immune response which is complex and strong enough to prevent infection of the wild-type virus. Live virus vaccines are typically more protective than single protein vaccines because it is harder to out-mutate T and B-cell responses to a larger number of epitopes. The weak growth of the live virus vaccine results in a larger antigenic dose and point mutation is increases the complexity of the immune response. To evaluate vaccine competence, vaccine potential is evaluated in Macaques (*M. nemestrina*) or chimpanzees using SIV variants that are known to cause AIDS. Sequence for an example SIV, SIVsmm, is found at Gene Bank Accession No. x14307. This virus is closely related to HIV-2. *See*, Hirsch (1989) Nature 339: 389-392. In general, many complete sequences for HIVs, SIVs and many other viruses

are found in well known sequence repositories, including GenBank, EMBL, DDBJ and the NCBI. Well characterized HIV clones include: HIV-1NL43, HIV-1SF2, HIV-1BRU and HIV-1MN. For an introduction to the genetic variability of HIV, *see,* Seillier-Moiseiwitsch et al. (1994) Annu. Rev. Genet. 28:559-96 and the references cited therein.

Several HIV-2 isolates, including three molecular clones of HIV-2 (HIV-2_{ROD}, HIV-$2_{SBL-ISY}$, and HIV-2_{UC1}), have also been reported to infect macaques (*M. mulatta* and *M. nemestrina*) or baboons (Franchini, *et al.* (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2433-2437; Barnett, *et al.* (1993) *Journal of Virology* 67, 1006-14; Boeri, *et al.* (1992) *Journal of Virology* 66, 4546-50; Castro, *et al.* (1991) *Virology* 184, 219-26; Franchini, *et al.* (1990) *Journal of Virology* 64, 4462-7; Putkonen, *et al.* (1990) *Aids* 4, 783-9; Putkonen, *et al.* (1991) *Nature* 352, 436-8). As human pathogens capable of infection of small primates, HIV-2 molecular clones provide attractive models for studies of AIDS pathogenesis, and for drug and vaccine development against HIV-1 and HIV-2.

Recently, HIV-2 was suggested as a possible vaccine candidate against the more virulent HIV-1 due to its long asymptomatic latency period, and its ability to protect against infection by HIV-1 (*see*, Travers *et al.* (1995) *Science* 268: 1612-1615 and related commentary by Cohen *et al* (1995) *Science* 268: 1566). In the nine-year study by Travers *et al.* (*id*) of West African prostitutes infected with HIV-2, it was determined that infection with HIV-2 caused a 70% reduction in infection by HIV-1. Thus, codon altered HIV-2 viruses can also be used as a live vaccine, against both HIV-2 and HIV-1. Furthermore, because the natural pathogenicity of HIV-2 is less than HIV-1, it is, in addition to HIV-1, a preferred virus for modification.

FORMATS FOR SEQUENCE RECOMBINATION

The methods of the invention entail performing recombination ("shuffling") and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer (1995) *Bio/Technology* 13:549-553). Reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest. Such techniques do not require the extensive analysis and computation required by conventional methods for polypeptide engineering. Shuffling allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to natural pair-wise recombination events (e.g., as occur during sexual

replication). Thus, the sequence recombination techniques described herein provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result. In some instances, however, structural and/or functional information is available which, although not required for sequence recombination, provides opportunities for modification of the technique.

Exemplary formats and examples for sequence recombination, referred to, e.g., as "DNA shuffling," "fast forced evolution," or "molecular breeding," have been described by the present inventors and co-workers in the following patents and patent applications: US Patent No. 5,605,793; PCT Application WO 95/22625 (Serial No. PCT/US95/02126), filed February 17, 1995; US Serial No. 08/425,684, filed April 18, 1995; US Serial No. 08/621,430, filed March 25, 1996; PCT Application WO 97/20078 (Serial No. PCT/US96/05480), filed April 18, 1996; PCT Application WO 97/35966, filed March 20, 1997; US Serial No. 08/675,502, filed July 3, 1996; US Serial No. 08/721, 824, filed September 27, 1996; PCT Application WO 98/13487, filed September 26, 1997; "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination" Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre et al. (PCT/US99/15972, filed 07/15/1999); Stemmer, *Science* 270:1510 (1995); Stemmer *et al.*, *Gene* 164:49-53 (1995); Stemmer, *Bio/Technology* 13:549-553 (1995); Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* 91:10747-10751 (1994); Stemmer, *Nature* 370:389-391 (1994); Crameri *et al.*, *Nature Medicine* 2(1):1-3 (1996); and Crameri *et al.*, *Nature Biotechnology* 14:315-319 (1996), each of which is incorporated by reference in its entirety for all purposes.

The recombination procedure starts with at least two substrates that generally show substantial sequence identity to each other (*e.g.*, at least about 30%, 50%, 70%, 80% or 90% or more sequence identity), but differ from each other at certain positions. For example, at least one codon altered nucleic acid is recombined with one or more additional nucleic acid (the additional nucleic acid can also be a codon altered nucleic acid) herein. The difference between nucleic acids to be recombined can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in about 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide

positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. More typically, they will be codon altered nucleic acids derived from one or more homologous nucleic acid sequence. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness (*e.g.*, codon altered nucleic acids derived from different, but homologous, genes within a superfamily). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example; one segment can be a subsequence of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form.

The starting DNA segments are recombined by any of the sequence recombination formats provided herein to generate a diverse library of recombinant DNA segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5, 10^9, 10^{12}, 10^{15}, 10^{20} or even more members. In some embodiments, the starting segments and the recombinant libraries generated will include essentially full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. In other embodiments, the recombinant DNA segments in the library can be inserted into a common vector providing sequences necessary for expression before performing screening/selection.

Use of Restriction Enzyme Sites to Recombine Mutations

In some situations it is advantageous to use restriction enzyme sites in nucleic acids to direct the recombination of mutations in a nucleic acid sequence of interest. These techniques are particularly preferred in the evolution of fragments that cannot readily be shuffled by other existing methods due to the presence of repeated DNA or other problematic

primary sequence motifs. These situations also include recombination formats in which it is preferred to retain certain sequences unmutated. The use of restriction enzyme sites is also preferred for shuffling large fragments (typically greater than 10 kb), such as gene clusters that cannot be readily shuffled and "PCR-amplified" because of their size. Although fragments up to 50 kb have been reported to be amplified by PCR (Barnes, *Proc. Natl. Acad. Sci. U.S.A.* 91:2216-2220 (1994)), it can be problematic for fragments over 10 kb, and thus alternative methods for shuffling in the range of 10 - 50 kb and beyond are preferred. Preferably, the restriction endonucleases used are of the Class II type (Sambrook, Ausubel and Berger, *supra*) and of these, preferably those which generate nonpalindromic sticky end overhangs such as Alwn I, Sfi I or BstX1. These enzymes generate nonpalindromic ends that allow for efficient ordered reassembly with DNA ligase. Typically, restriction enzyme (or endonuclease) sites are identified by conventional restriction enzyme mapping techniques (Sambrook, Ausubel, and Berger, *supra.*), by analysis of sequence information for that gene, or by introduction of desired restriction sites into a nucleic acid sequence by synthesis (*i.e.* by incorporation of silent mutations). For example, one or more codon-altered nucleic acid can be recombined at restriction sites, e.g., with one or more nucleic acid of interest (including, e.g. a gene or gene cluster to be modified by recombination with the codon-altered nucleic acid).

The DNA substrate molecules to be digested can either be from *in vivo* replicated DNA, such as a plasmid preparation, or from synthetic or e.g., PCR amplified nucleic acid fragments harboring the restriction enzyme recognition sites of interest, preferably near the ends of the fragment. Typically, at least two variants of a gene of interest, each having one or more mutations, and at least one of which incorporating codon-modifications, are digested with at least one restriction enzyme determined to cut within the nucleic acid sequence of interest. The restriction fragments are then joined with DNA ligase to generate full length genes having shuffled regions. The number of regions shuffled will depend on the number of cuts within the nucleic acid sequence of interest. The shuffled molecules can be introduced into cells as described above and screened or selected for a desired property as described herein. Nucleic acid can then be isolated from pools (libraries), or clones having desired properties and subjected to the same procedure until a desired degree of improvement is obtained.

In some embodiments, at least one DNA substrate molecule or fragment thereof is isolated and subjected to mutagenesis. In some embodiments, the pool or library of religated restriction fragments are subjected to mutagenesis or additional recombination protocols before the digestion-ligation process is repeated. "Mutagenesis" as used herein comprises such techniques known in the art as PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, etc., and recursive sequence recombination by any of the techniques described herein.

Reassembly PCR

A further technique for recombining mutations in a nucleic acid sequence utilizes "reassembly PCR." This method can be used to assemble multiple segments that have been separately evolved into a full length nucleic acid template such as a gene. This technique is performed when a pool of advantageous mutants is known from previous work or has been identified by screening mutants that may have been created by any mutagenesis technique known in the art, such as PCR mutagenesis, cassette mutagenesis, doped oligo mutagenesis, chemical mutagenesis, or propagation of the DNA template *in vivo* in mutator strains. Boundaries defining segments of a nucleic acid sequence of interest preferably lie in intergenic regions, introns, or areas of a gene not likely to have mutations of interest. Preferably, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of the nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length.

Each of the segments is amplified with a set of such primers. The PCR products are then "reassembled" according to assembly protocols such as those discussed herein to assemble randomly fragmented genes. In brief, in an assembly protocol the PCR products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTPs) and appropriate buffer salts in the absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes. In some embodiments, the resulting reassembled genes are subjected to mutagenesis before the process is repeated.

In the present invention, oligos such as PCR primers can include codon modifications as compared to a starting sequence. In addition, oligonucleotides can form the basis for PCR concatemerization reactions in which overlapping hybridized oligonucleotides are extended in one or more PCR amplification cycles. In this embodiment, a template nucleic acid is not required (although a template or fragments thereof can be added to the amplification mixture, which can aid in the eventual reassembly of a full-length gene). Further details regarding oligonucleotide gene reassembly methods are found, e.g., in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049.

In a further embodiment, the PCR primers for amplification of segments of a nucleic acid sequence of interest are used to introduce variation into the gene of interest as follows. Mutations at sites of interest in a nucleic acid sequence are identified by screening or selection, by sequencing homologues of the nucleic acid sequence, and so on. Oligonucleotide PCR primers are synthesized which encode wild type or mutant information at sites of interest. These primers are then used in PCR mutagenesis to generate libraries of full length genes encoding permutations of wild type and mutant information at the designated positions. This technique is typically advantageous in cases where the screening or selection process is expensive, cumbersome, or impractical relative to the cost of sequencing the genes of mutants of interest and synthesizing mutagenic oligonucleotides.

Site Directed Mutagenesis (SDM) with Oligonucleotides Encoding Homologue Mutations Followed by Shuffling

In some embodiments of the invention, sequence information from one or more substrate sequences is added to a given "parental" sequence of interest, with subsequent recombination between rounds of screening or selection. Typically, this is done with site-directed mutagenesis performed by techniques well known in the art (e.g., Berger, Ausubel and Sambrook, *supra.*) with one substrate as a template and oligonucleotides encoding single or multiple mutations from other substrate sequences, *e.g.* homologous genes. After screening or selection for an improved phenotype of interest, the selected recombinant(s) can be further evolved using recursive techniques. After screening or selection, site-directed mutagenesis can be done again with another collection of oligonucleotides encoding

homologue mutations, and the above process repeated until the desired properties are obtained.

When the difference between two homologues is one or more single point mutations in a codon, degenerate oligonucleotides can be used that encode the sequences in both homologues. One oligonucleotide can include many such degenerate codons and still allow one to exhaustively search all permutations over that block of sequence.

When the homologue sequence space is very large, it can be advantageous to restrict the search to certain variants. Thus, for example, computer modeling tools (Lathrop *et al.* (1996) *J. Mol. Biol.*, 255: 641-665) can be used to model each homologue mutation onto the target protein and discard any mutations that are predicted to grossly disrupt structure and function. In silico genetic algorithm operations for generating and predicting mutational events are found in Selifonov and Stemmer "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

Oligonucleotide and in silico shuffling formats

As mentioned above, at least two additional related formats are useful in the practice of the present invention. The first, referred to as "in silico" shuffling utilizes computer algorithms to perform "virtual" shuffling using genetic operators in a computer. As applied to the present invention, codon altered gene sequence strings are recombined in a computer system and desirable products are made, e.g., by reassembly PCR or ligation of synthetic oligonucleotides. In silico shuffling is described in detail in Selifonov and Stemmer in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854. In brief, genetic operators (algorithms which represent given genetic events such as point mutations, recombination of two strands of homologous nucleic acids, etc.) are used to model recombinational or mutational events which can occur in one or more nucleic acid, e.g., by aligning nucleic acid sequence strings (using standard alignment software, or by manual inspection and alignment) and predicting recombinational outcomes. The predicted recombinational outcomes are used to produce corresponding molecules, e.g., by oligonucleotide synthesis and reassembly PCR.

The second useful format is referred to as "oligonucleotide mediated shuffling" in which oligonucleotides corresponding to a family of related homologous nucleic acids (e.g., as applied to the present invention, codon modified synthetic homologous variants of a nucleic acid) which are recombined to produce selectable nucleic acids. This format is described in detail in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. In brief, selected oligonucleotides are synthesized, ligated and elongated, typically either in a polymerase or ligase-mediated elongation reaction. The technique can be used to recombine homologous or even non-homologous codon-altered nucleic acid sequences.

One advantage of oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of fragmented nucleic acids (e.g., cleaved codon-modified oligonucleotides, or synthesized codon-modified oligonucleotides) are recombined, e.g., with a with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination.

When recombining homologous nucleic acids, sets of overlapping family gene shuffling oligonucleotides (which are derived by comparison of homologous nucleic acids that include one or more codon-modified nucleic acid, followed by synthesis of corresponding oligonucleotides) are hybridized and elongated (e.g., by reassembly PCR or ligation), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. The set of overlapping family shuffling gene oligonucleotides includes a plurality of oligonucleotide member types which have consensus region subsequences derived from a plurality of homologous target nucleic acids.

Typically, as applied to the present invention, family gene shuffling oligonucleotide which include one or more codon-altered nucleic acid(s) are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are synthesized (serially or in parallel) which correspond to at least one region of sequence diversity.

Sets of fragments, or subsets of fragments used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (e.g., with a DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments can be used in conjunction with family gene shuffling oligonucleotides, e.g., in one or more recombination reaction to produce recombinant codon-altered nucleic acid(s).

Additional oligonucleotide shuffling formats are found in co-filed application by Crameri et al., "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" (Attorney Docket Number 02-296-2US) and in co-filed application by Welch et al., "USE OF CODON VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" (Attorney docket number 02-1007). In particular, these applications provide for tri-nucleotide-based synthesis of degenerate oligonucleotides, thereby providing for codon substitution during oligonucleotide shuffling. In brief, this procedure utilizes tri-nucleotide phosphoramidite chemistry to synthesize oligos, rather than standard mono-nucleotide synthesis. Because codons are altered as a unit, the synthetic scheme of degenerate oligonucleotides is simplified.

Additional In Vitro DNA Shuffling Formats

In one embodiment for shuffling DNA sequences *in vitro*, the initial substrates for recombination are a pool of related sequences, *e.g.*, different variant forms, as homologs from different individuals, strains, or species of an organism, or related sequences from the same organism, as allelic variations. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or reassembled. Preferably the sequences are from 50 base pairs (bp) to 50 kilobases (kb).

The pool of related substrates are converted into overlapping fragments, *e.g.*, from about 5 bp to 5 kb or more. Often, for example, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNase I or RNase digestion, random shearing or partial restriction enzyme digestion, or by oligonucleotide synthesis as in the family oligonucleotide-mediated shuffling methods of crameri et al., discusses *supra*. For discussions of protocols for the isolation, manipulation, enzymatic digestion, and the like, of nucleic acids, *see*, for example, Sambrook *et al.* and Ausubel, both *supra*. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 2, 10, 100, 500 or 1,000 or more.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form using any of a variety of techniques, including, for example, heating, chemical denaturation, use of DNA binding proteins, and the like (in oligonucleotide mediated methods, this step can be omitted). Conversion can be effected by heating to about 80 °C to 100 °C, more preferably from 90 °C to 96 °C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with a single-stranded DNA binding protein (see Wold (1997) *Annu. Rev. Biochem.* 66:61-92) or *recA* protein (*see, e.g.*, Kiianitsa (1997) *Proc. Natl. Acad. Sci. U S A* 94:7837-7840). Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 20 °C to 75 °C, and preferably from 40 °C to 65 °C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can

be used with an annealing temperature of between 20-30 °C. The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase or ligase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as shuffling of the nucleic acid *in vitro*. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from shuffling is used to transform host cells, optionally after cloning into a vector.

In one embodiment utilizing *in vitro* shuffling, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. Another embodiment uses random primers to prime an entire template DNA to generate less than full length amplification products. The amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, in which at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the partially extended (less than full length) products reanneal to and prime extension on different sequence-related template species. In another embodiment, the conversion of substrates to fragments can be effected by partial PCR amplification of substrates.

In another embodiment, a mixture of fragments is spiked with one or more oligonucleotides. The oligonucleotides can be designed to include precharacterized mutations of a wildtype sequence (e.g., codon modification), or sites of natural variations between individuals or species. The oligonucleotides also typically include sufficient sequence or structural homology flanking such mutations or variations to allow annealing

with the wildtype fragments. Annealing temperatures can be adjusted depending on the length of homology.

In a further embodiment, recombination occurs in at least one cycle by template switching, such as when a DNA fragment derived from one template primes on the homologous position of a related but different template. Template switching can be induced by addition of recA (*see*, Kiianitsa (1997) *supra*), rad51 (*see*, Namsaraev (1997) *Mol. Cell. Biol.* 17:5359-5368), rad55 (*see*, Clever (1997) *EMBO J.* 16:2535-2544), rad57 (*see*, Sung (1997) *Genes Dev.* 11:1111-1121) or other polymerases (*e.g.*, viral polymerases, reverse transcriptase) to the amplification mixture. Template switching can also be increased by increasing the DNA template concentration.

Another embodiment utilizes at least one cycle of amplification, which can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Fragments can be prepared using a single stranded DNA phage, such as M13 (*see*, Wang (1997) *Biochemistry* 36:9486-9492). Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, ssDNA fragments of variable length can be generated from a single primer by Pfu, Taq, Vent, Deep Vent, UlTma DNA polymerase or other DNA polymerases on a first DNA template (*see*, Cline (1996) *Nucleic Acids Res.* 24:3546-3551). The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular ssDNA. This results in multiple substitutions of the first template into the second. *See*, Levichkin (1995) *Mol. Biology* 29:572-577; Jung (1992) *Gene* 121:17-24.

In some embodiments of the invention, shuffled nucleic acids obtained by use of the recursive recombination methods of the invention, are put into a cell and/or organism for screening. Shuffled genes can be introduced into, for example, bacterial cells, yeast cells, fungal cells vertebrate cells, invertebrate cells or plant cells for initial screening. *Bacillus* species (such as *B. subtilis* and *E. coli* are two examples of suitable bacterial cells into which one can insert and express shuffled genes which provide for convenient shuttling to other cell types (a variety of vectors for shuttling material between these bacterial cells and eukaryotic cells are available; *see*, Sambrook, Ausubel and Berger, *all supra*). The shuffled genes can

be introduced into bacterial, fungal or yeast cells either by integration into the chromosomal DNA or as plasmids.

Bacterial, plant, animal and yeast systems are preferred in the present invention. For example, in one embodiment, shuffled genes can be introduced into plant or animal cells for production purposes (it will be appreciated that transgenic plants are, increasingly, an important source of industrial enzymes), or can be introduced into a plant or animal cell for therapeutic purposes. Thus, a transgene of interest can be modified using the recursive sequence recombination methods of the invention *in vitro* and reinserted into the cell for *in vivo/in situ* selection for the new or improved property, in bacteria, eukaryotic cells, or whole eukaryotic organisms.

In Vivo DNA Shuffling Formats

In some embodiments of the invention, DNA substrate molecules, e.g., those comprising codon modifications relative to a wild-type sequence, are introduced into cells, where the cellular machinery directs their recombination. For example, a library of mutants is constructed and screened or selected for mutants with improved phenotypes by any of the techniques described herein.

The DNA substrate molecules encoding the best candidates are recovered by any of the techniques described herein, then fragmented and used to transfect a plant host and screened or selected for improved function. If further improvement is desired, the DNA substrate molecules are recovered from the host cell, such as by PCR, and the process is repeated until a desired level of improvement is obtained. In some embodiments, the fragments are denatured and reannealed prior to transfection, coated with recombination stimulating proteins such as recA, or co-transfected with a selectable marker such as Neo^R to allow the positive selection for cells receiving recombined versions of the gene of interest. Methods for *in vivo* shuffling are described in, for example, PCT application WO 98/13487 and WO 97/20078. The efficiency of *in vivo* shuffling can be enhanced by increasing the copy number of a gene of interest in the host cells.

Whole Genome Shuffling

In one embodiment, the selection methods herein are utilized in a "whole genome shuffling" format. An extensive guide to the many forms of whole genome shuffling is found in the pioneering application to the inventors and their co-workers entitled

"Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," PCT/US99/15972, by del Cardayre et al. Any codon-altered set of nucleic acids can be used to transform cells, which can then be shuffled by in a whole genome format.

In brief, whole genome shuffling makes no presuppositions at all regarding what nucleic acids may confer a desired property. Instead, entire genomes (e.g., from a genomic -library, or isolated from an organism) are shuffled in cells and selection protocols applied to the cells. These genomes can be spiked with any desired set of nucleic acids, including codon-modified nucleic acids.

Assays

The relevant assay for selection of a desired property of a codon-modified nucleic acid will depend on the application. Many assays which detect activity for proteins, receptors, ligands, cells and the like are known. Formats include binding to immobilized components, cell or organismal viability, production of reporter compositions, and the like.

In the high throughput assays of the invention, it is possible to screen up to several thousand different shuffled variants in a single day. In particular, each well of a microtiter plate can be used to run a separate assay, or, if concentration or incubation time effects are to be observed, every 5-10 wells can test a single variant. Thus, a single standard microtiter plate can assay about 100 (e.g., 96) reactions. If 1536 well plates are used, then a single plate can easily assay from about 100- about 1500 different reactions. It is possible to assay several different plates per day; assay screens for up to about 6,000-20,000 different assays (i.e., involving different nucleic acids, encoded proteins, concentrations, etc.) is possible using the integrated systems of the invention. More recently, microfluidic approaches to reagent manipulation have been developed, e.g., by Caliper Technologies (Mountain View, CA).

In one aspect, library members, e.g., cells, viral plaques, spores or the like, are separated on solid media to produce individual colonies (or plaques). Using an automated colony picker (e.g., the Q-bot, Genetix, U.K.), colonies or plaques are identified, picked, and up to 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or viruses in plaque applications). The time the pin is in the colony, the number of dips to

inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each can be controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. The glass balls in the microtiter plates act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter. Clones from cultures of interest can be cloned by limiting dilution. As also described supra, plaques or cells constituting libraries can also be screened directly for production of proteins, either by detecting hybridization, protein activity, protein binding to antibodies, or the like.

The ability to detect a subtle increase in the performance of a shuffled library member over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size, a prescreen that increases the number of mutants processed by, e.g., 10-fold can be used. The goal of the primary screen is to quickly identify mutants having equal or better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture for subsequent analysis.

A number of well known robotic systems have also been developed for solution phase chemistries useful in assay systems. These systems include automated workstations like the automated synthesis apparatus developed by Takeda Chemical Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Hewlett-Packard, Palo Alto, Calif.) which mimic the manual synthetic operations performed by a scientist. Any of the above devices are suitable for use with the present invention, e.g., for high-throughput screening of molecules encoded by codon-altered nucleic acids. The nature and implementation of modifications to these devices (if any) so that they can operate as discussed herein with reference to the integrated system will be apparent to persons skilled in the relevant art.

High throughput screening systems are commercially available (*see, e.g.,* Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures including all sample and reagent pipetting, liquid

dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization.

The manufacturers of such systems provide detailed protocols the various high throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like. Microfluidic approaches to reagent manipulation have also been developed, e.g., by Caliper Technologies (Mountain View, CA).

Optical images viewed (and, optionally, recorded) by a camera or other recording device (*e.g.*, a photodiode and data storage device) are optionally further processed in any of the embodiments herein, *e.g.*, by digitizing the image and/or storing and analyzing the image on a computer. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, *e.g.*, using PC (Intel x86 or pentium chip- compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (*e.g.*, SUN™ work station) computers.

One conventional system carries light from the assay device to a cooled charge-coupled device (CCD) camera, in common use in the art. A CCD camera includes an array of picture elements (pixels). The light from the specimen is imaged on the CCD. Particular pixels corresponding to regions of the specimen (*e.g.*, individual hybridization sites on an array of biological polymers) are sampled to obtain light intensity readings for each position. Multiple pixels are processed in parallel to increase speed. The apparatus and methods of the invention are easily used for viewing any sample, e.g., by fluorescent or dark field microscopic techniques.

Software elements for manipulating strings of characters which correspond to codon-modified nucleic acids can be used to direct synthesis of oligonucleotides relevant to shuffling of codon-modified nucleic acids. Integrated systems comprising these and other useful features, e.g., a digital computer with additional features such as high-throughput liquid control software, image analysis software, data interpretation software, a robotic liquid control armature for transferring solutions from a source to a destination operably linked to the digital computer, an input device (e.g., a computer keyboard) for entering data to the

digital computer to control high throughput liquid transfer by the robotic liquid control armature an image scanner for digitizing label signals from labeled assay components, or the like are a feature of the invention.

In one aspect, the invention provides an integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database. As discussed theroughout, there are a variety of sequence database programs for aligning and manipulating sequences. In addition, standard text manipulation software such as word processing software (e.g., Microsft Word™ or Corel Wodperfect™) and database software (e.g., spreadsheet software such as Microsoft Excel™, Corel Quattro Pro™, or database programs such as Microsoft Access™ or Paradox™) can be used in conjuction with a user interface (e.g., a GUI in a standard operating system such as a Windows, Macintosh or LINUX system) to manipulate strings of characters. Specialized alignment programs such as BLAST can also be incorporated into the systems of the invention for alignment of codon-altered nucleic acids (or corresponding character strings).

In addition to the integrated system elements mentioned above, the integrated system can also include an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium. Typically, the synthesizer is programmed to synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to test shuffled codon-modified DNAs, including in an iterative process.

An assay, kit or system utilizing a use of any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits will optionally additionally comprise instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally comprise one or more of the following: (1) a shuffled codon-modified component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein; (3) one or more assay component; (4) a container for holding nucleic acids or enzymes, other nucleic acids, transgneic plants, animals, cells, or the like, (5) packaging materials and (6) software fixerd in a computer readable medium comprising sequences corresponding to one or more codon-altered nucleic acid character string.

In a further aspect, the present invention provides for the use of any component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and materials described above can be used in various combinations. All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

WHAT IS CLAIMED IS:

1. A method of making codon altered nucleic acids, the method comprising:
(i) providing a first nucleic acid sequence, which nucleic acid sequence encodes a first polypeptide sequence;
(ii) providing a plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide or a modified form thereof; and,
(iii) recombining the plurality of codon-altered nucleic acid sequences to produce a target codon altered nucleic acid, which target codon altered nucleic acid encodes a second protein.

2. The method of claim 1, wherein at least one of the plurality of codon altered nucleic acid sequences does not hybridize to the first nucleic acid under stringent hybridization conditions.

3. The method of claim 1, further comprising shuffling a nucleic acid comprising a subsequence consisting of the first nucleic acid, or a substantially identical variant thereof, with one or more of the plurality of codon altered nucleic acids, or with the target codon altered nucleic acid.

4. The method of claim 1, the method further comprising the step of:
(iv) screening the second protein for a structural or functional property.

5. The method of claim 1, the method further comprising the steps of:
(iv) screening the second protein for a structural or functional property, and,
(v) comparing the structural or functional property of the second protein to a structural or functional property of the first protein.

6. The method of claim 1, wherein the second polypeptide has a structural or functional property equivalent or superior to the first polypeptide.

7. The method of claim 1, wherein the first and second polypeptide are homologous.

8. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered nucleic acids.

9. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered conservatively modified nucleic acids.

10. A library of codon altered conservatively modified nucleic acids produced by the method of claim 9.

11. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered non-conservatively modified nucleic acids.

12. A library of codon altered conservatively modified nucleic acids produced by the method of claim 11.

13. The method of claim 1, wherein the plurality of codon altered nucleic acids is derived from a plurality of forms of the first nucleic acid.

14. The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise at least three codon altered nucleic acids.

15. The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise one or more of the following structural features:

(a) codon usage divergence for each of the codon altered nucleic acids of 50% or more as compared to the first nucleic acid;

(b) codon usage divergence for each of the codon altered nucleic acids of 75% or more as compared to the first nucleic acid;

(c) codon usage divergence for each of the codon altered nucleic acids of 90% or more as compared to the first nucleic acid;

(d) maximal codon usage divergence for each of the codon altered nucleic acids as compared to the first nucleic acid;

(e) non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid;

(f) a lack of high stringency hybridization between one or more of the codon altered nucleic acid and the first nucleic acid; and,

(g) modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide.

16. The method of claim 1, wherein the percent identity between the second protein and the first protein is lower than the percent identity between two of the plurality of codon altered nucleic acids.

17. The method of claim 1, wherein the first nucleic acid encodes a protein selected from: EPO, G-CSF, a viral envelope protein, a cytokine, and a phosphatase.

18. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are isolated nucleic acids.

19. The target codon altered nucleic acid produced by the method of claim 1.

20. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are nucleic acids present in cells.

21. The cells produced by the method of claim 20.

22. The method of claim 1, wherein each of the codon altered nucleic acid sequences comprises at least two nucleotide differences when compared to the first nucleic acid.

23. The method of claim 1, further comprising introducing the target codon altered nucleic acid into a cell, or into a vector or virus.

24. The cell, vector or virus produced by the method of claim 23.

25. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus.

26. The attenuated virus produced by the method of claim 25.

27. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus, which attenuated virus produces an immune response upon infection by the virus in a mammal.

28. The attenuated virus produced by the method of claim 27.

29. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a retroviral genome to produce an attenuated retrovirus, which attenuated retrovirus produces an immune response upon infection by the retrovirus in a mammal.

30. The attenuated retrovirus produced by the method of claim 29.

31. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector.

32. The viral vector produced by the method of claim 31.

33. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector, which vector requires *trans* complementation for replication, and which vector has a reduced rate of reversion to a replicative form as compared to a corresponding viral vector which lacks a subsequence corresponding to the target codon altered nucleic acid.

34. The viral vector produced by the method of claim 33.

35. The method of claim 33, wherein the vector comprises viral elements from one or more of: a lentivirus, an adenovirus, a herpes virus, and an adeno-associated virus.

36. A method of making a library of codon-altered nucleic acids, the method comprising:

(i) selecting a first nucleic acid sequence, which nucleic acid sequence encodes a first polypeptide sequence; and,

(ii) making a plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide or a modified form thereof, wherein the plurality of codon altered nucleic acids comprise the library.

37. A codon-altered library made by the method of claim 36.

38. The library of claim 37, wherein said library comprises at least 2 codon altered nucleic acids.

39. The library of claim 37, wherein said library comprises at least 5 codon altered nucleic acids.

40. The library of claim 37, wherein said library comprises at least 10 codon altered nucleic acids.

41. The library of claim 37, wherein said library comprises at least 100 codon altered nucleic acids.

42. A kit comprising the library of claim 25 and one or more of: a container and instructional materials providing method step instructions for recombining two or more of members of the library.

43. The method of claim 41, further comprising recombining said plurality of codon altered nucleic acids to produce a shuffled codon-altered library.

44. The codon altered library made by the method of claim 43.

45. The method of claim 36, wherein said nucleic acids encode a protein selected from EPO, a cytokine, a phosphatase, and a viral envelope protein.

46. A composition comprising a plurality of codon altered nucleic acids, each of which encode a first polypeptide or a modified form thereof.

47. A library of codon-altered nucleic acids, comprising a plurality of codon-altered nucleic acids derived from a plurality of homologous nucleic acids.

48. The library of claim 47, wherein said plurality of codon altered nucleic acids recombine in vitro at an increased rate compared to said plurality of homologous nucleic acids.

49. The library of claim 47, wherein the level of identity among said plurality of codon-altered nucleic acids is at least as high as among a plurality of polypeptides encoded by said plurality of homologous nucleic acids.

50. An integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database.

51. The integrated system of claim 50, further comprising an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium, which synthesizer is programmed to synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

TPO - DNA (Monkey)

FIG.5A

Translation of Monkey EPO cDNA

SauJA
GATCCCCCCCCCCCCCTGGACAGCCCCTCTCCTCAGCCCCTGAACCTGGCCCTGCCC

CGCTGAACTTCCCCGGGATCAGGACTCCCGGTGGTCACCCCCCCCCTAGGTCCCTGAG

-27 -20
Met Gly Val His Glu Cys Pro Ala Trp
GGACCCCGGCCAGGCCCCGAGATC GGG GTG CAC GAA TGT CCT GCC TGG

-10
Leu Trp Leu Leu Leu Ser Leu Val Ser Leu Pro Leu Gly Leu Pro
CTG TGG CTT CTC CTG TCT CTC GTC TCC CTG CCC CTC GGC CTG CCA

-1 +1 10
Val Pro Gly Ala Pro Pro Arg Leu Ile Cys Asp Ser Arg Val Leu
CTC CCG GGC GCC CCA CCA CGC CTC ATC TGT GAC AGC CGA GTC CTG

20
Glu Arg Tyr Leu Leu Glu Ala Lys Glu Ala Glu Asn Val Thr Met
GAG AGG TAC CTC TTG GAG GCC AAG GAG GCC GAG AAT GTC ACG ATG

30 40
Gly Cys Ser Glu Ser Cys Asn Leu Asn Glu Asn Ile Thr Val Pro
GGC TGT TCC GAA AGC TGC AAC CTG AAT GAG AAT ATC ACC GTC CCA

Fig. 1

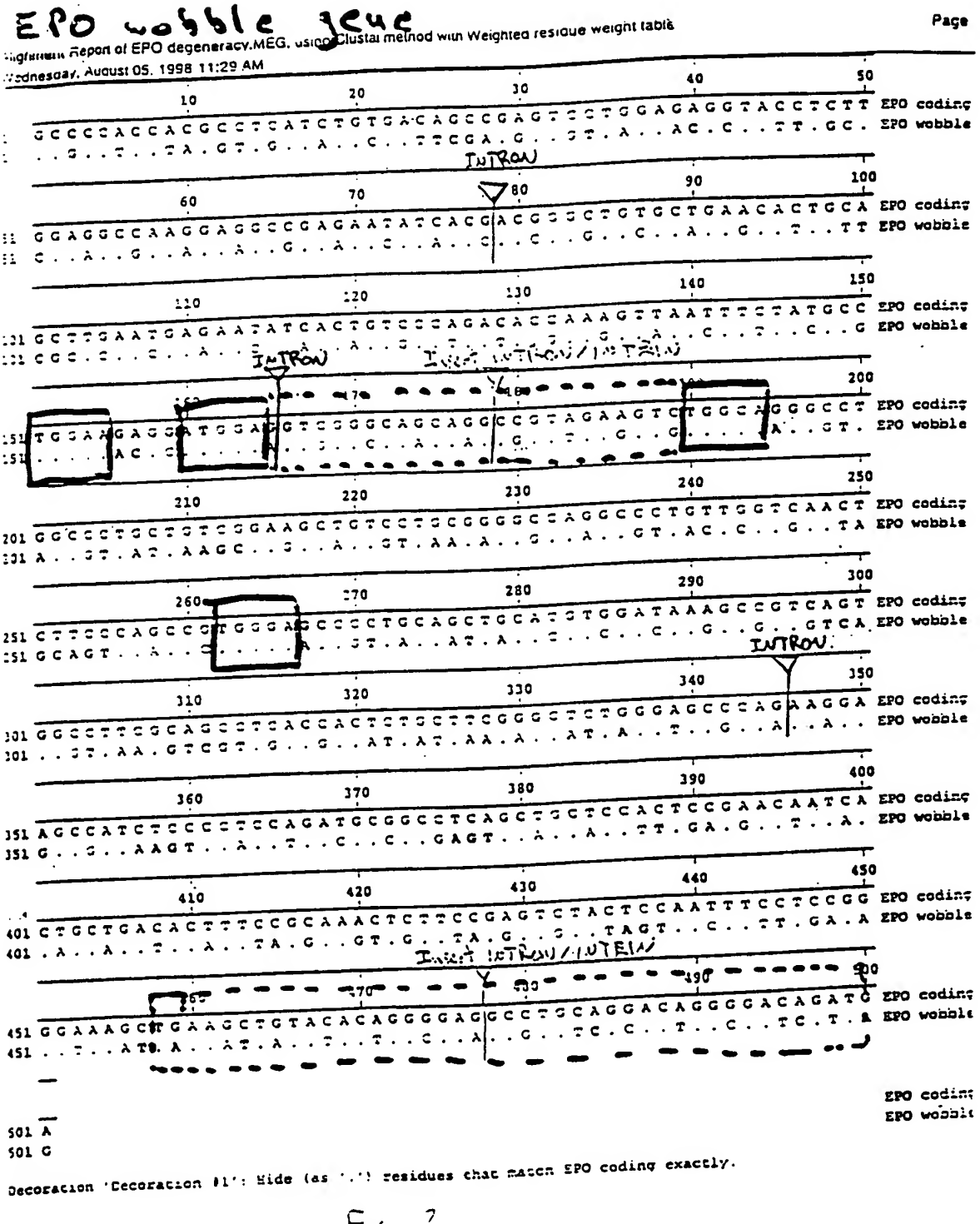

EPO wobble gene
Report of EPO degeneracy.MEG. using Clustal method with Weighted residue weight table
Page
Wednesday, August 05, 1998 11:29 AM
EPO coding
EPO wobble
INTRON
Decoration 'Decoration #1': Hide (as '.') residues that match EPO coding exactly.

Fig. 2

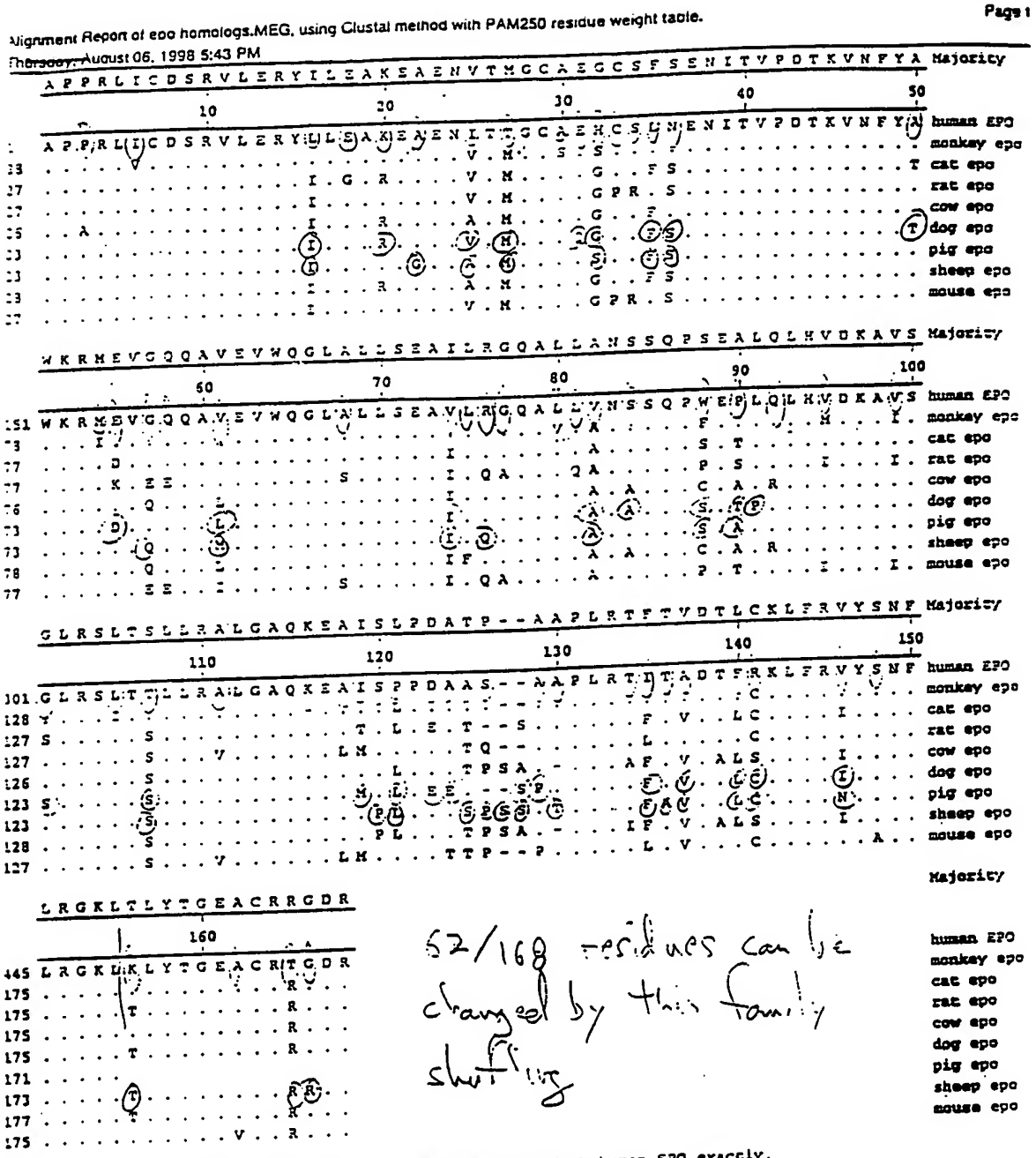

Alignment Report of epo homologs.MEG, using Clustal method with PAM250 residue weight table.
Page 1
Thursday, August 06, 1998 5:43 PM
Majority
human EPO
monkey epo
cat epo
rat epo
cow epo
dog epo
pig epo
sheep epo
mouse epo
52/168 residues can be changed by this family shifting
region of relative conservation
Decoration 'Decoration #1': Hide (as '.') residues that match human EPO exactly.

Fig. 3

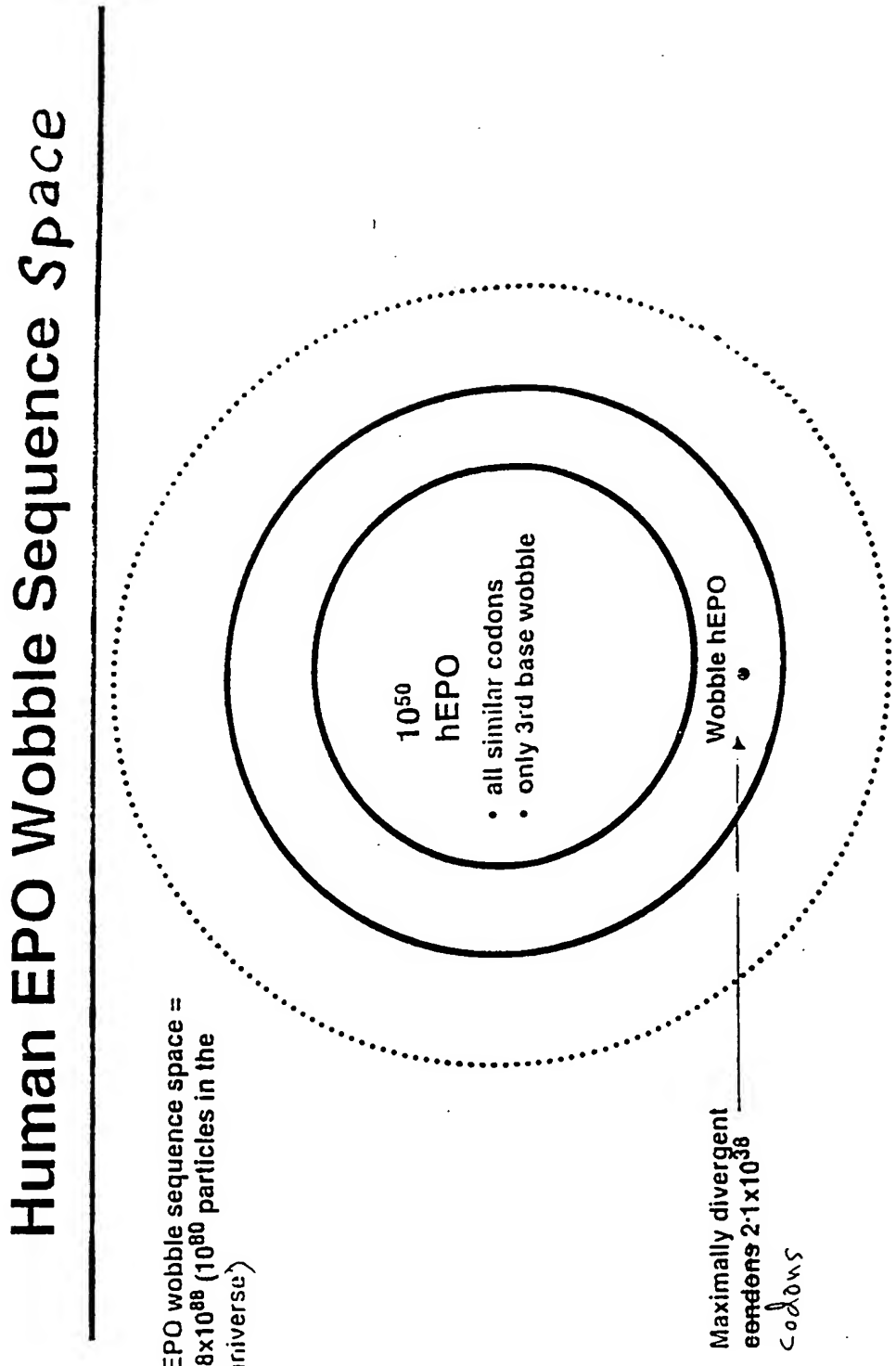
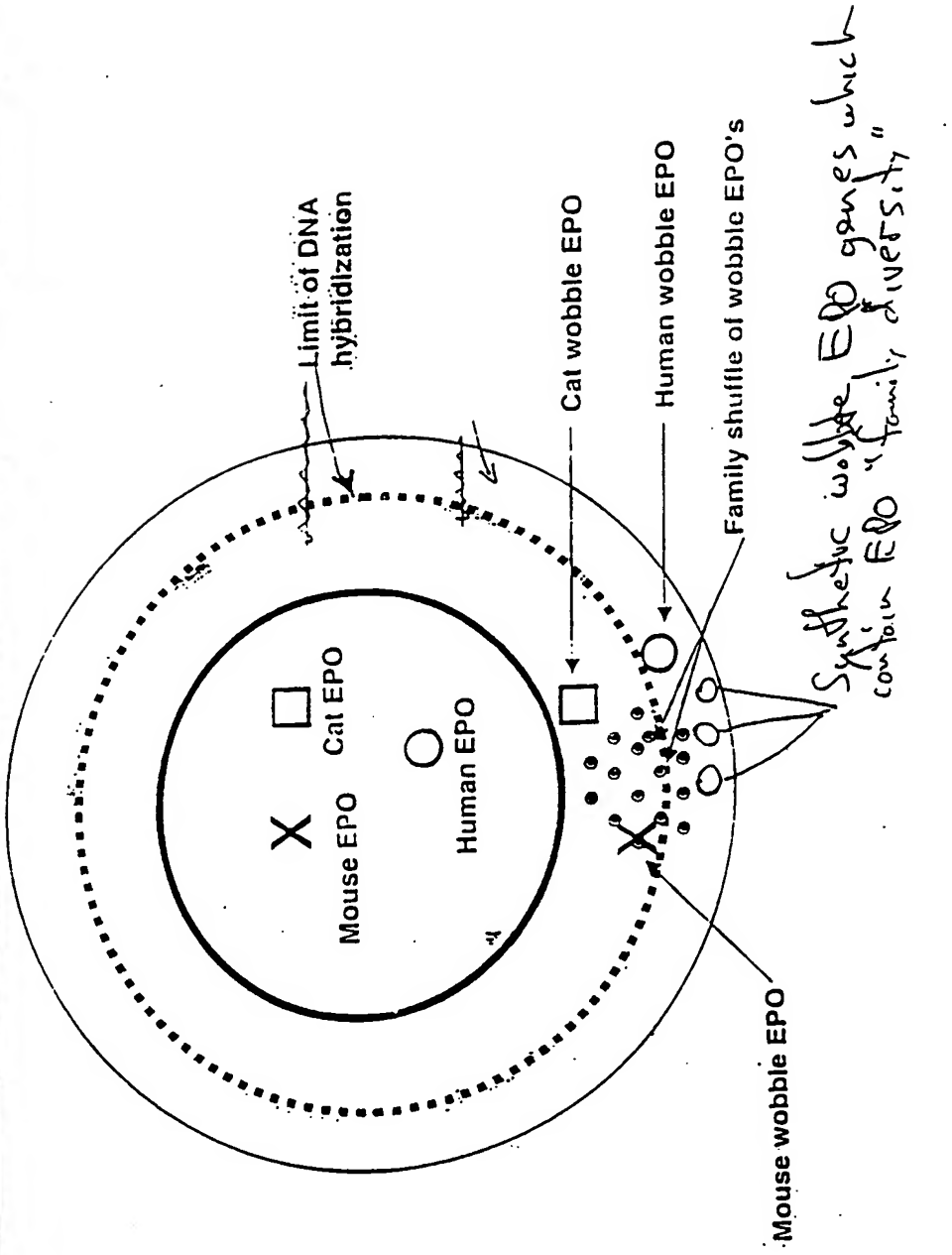

Human EPO Wobble Sequence Space
hEPO wobble sequence space = 2·8x10^88 (10^80 particles in the universe)
10^50
hEPO
• all similar codons
• only 3rd base wobble
Wobble hEPO
Maximally divergent ~~genes~~ 2·1x10^38 codons

Fig. 4

Fig. 5


Mammalian EPO Family-Wobble Sequence Space
Limit of DNA hybridization
Cat wobble EPO
Human wobble EPO
Family shuffle of wobble EPO's
Cat EPO
Mouse EPO
Human EPO
Mouse wobble EPO
Synthetic wobble EPO genes which contain EPO "family diversity"

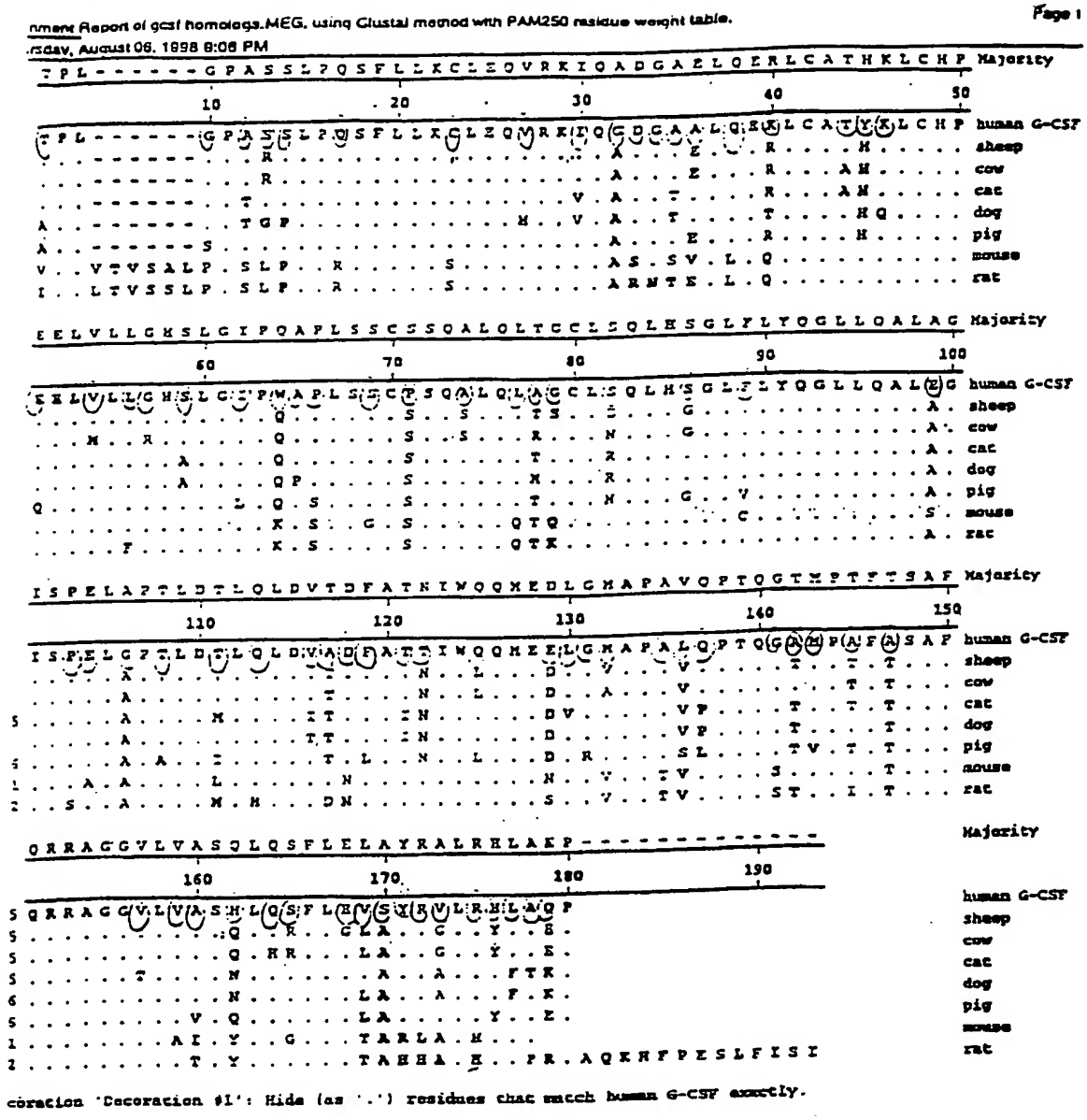

nment Report of gcsf homologs.MEG, using Clustal method with PAM250 residue weight table.
Page 1
nday, August 06, 1998 8:06 PM
Majority
human G-CSF
sheep
cow
cat
dog
pig
mouse
rat
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
coration 'Decoration #1': Hide (as '.') residues that match human G-CSF exactly.

Fig. 6

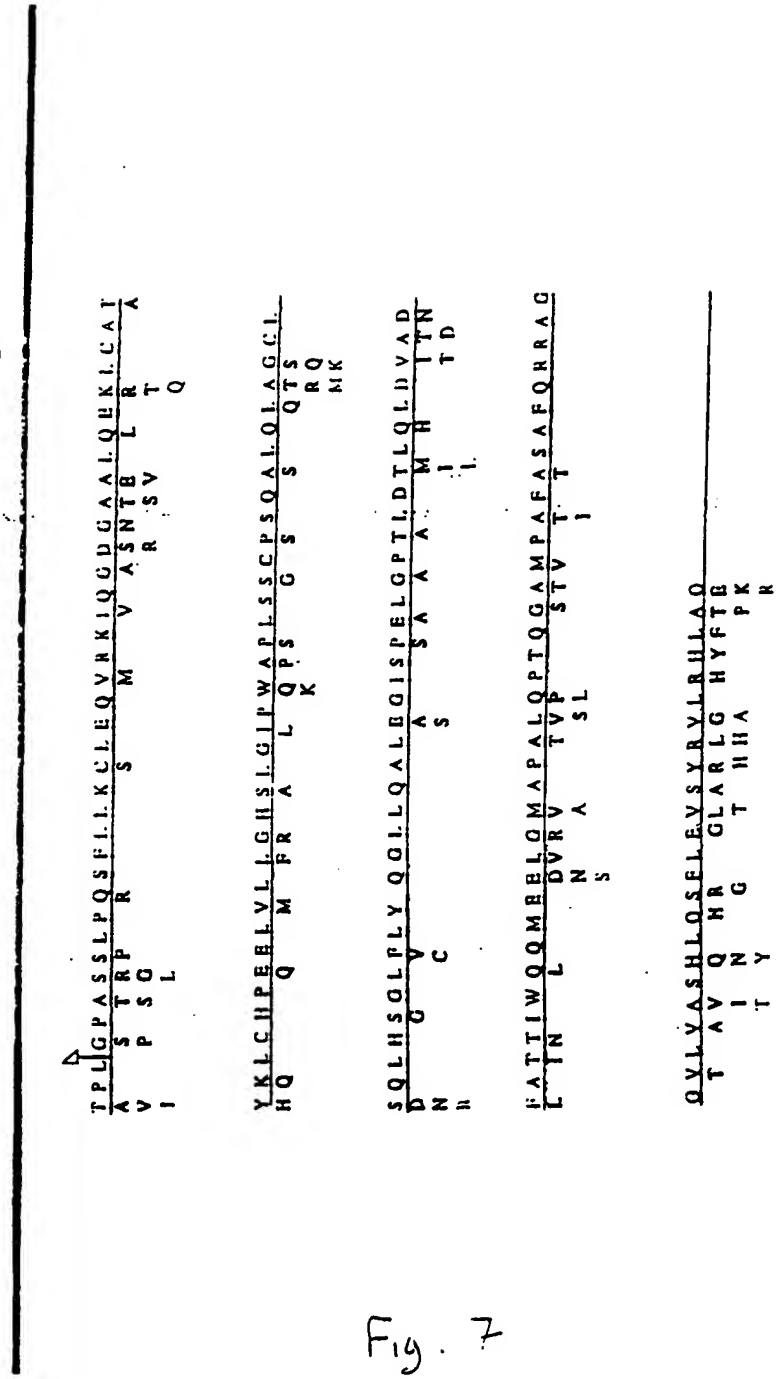
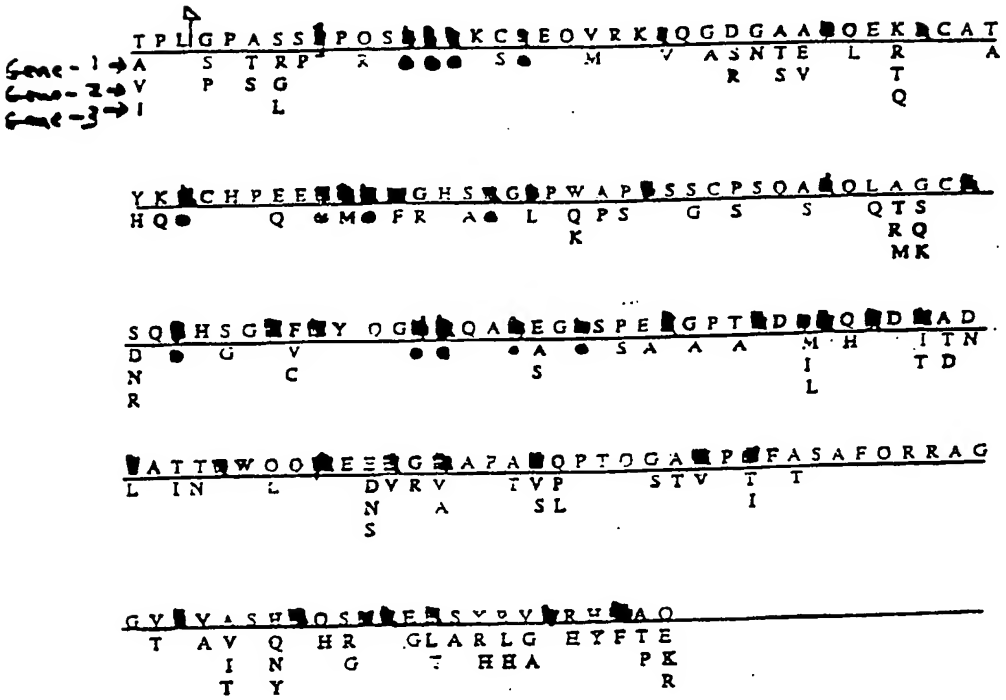


Fig. 7

Fig. 8

G-CSF

~~G-SSG~~ Family Diversity



• = hydrophobic core residues in conserved stretches of the protein

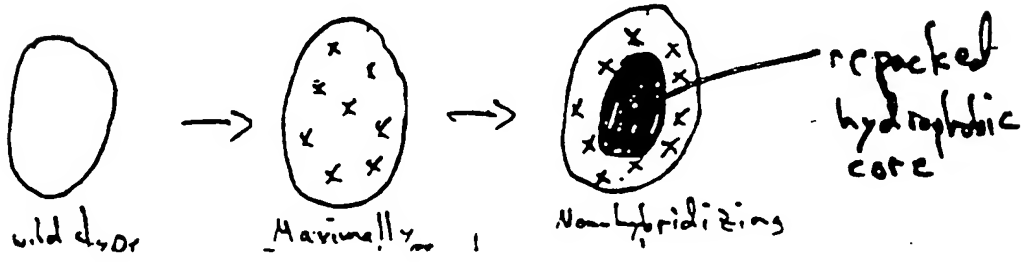


Fig. 9

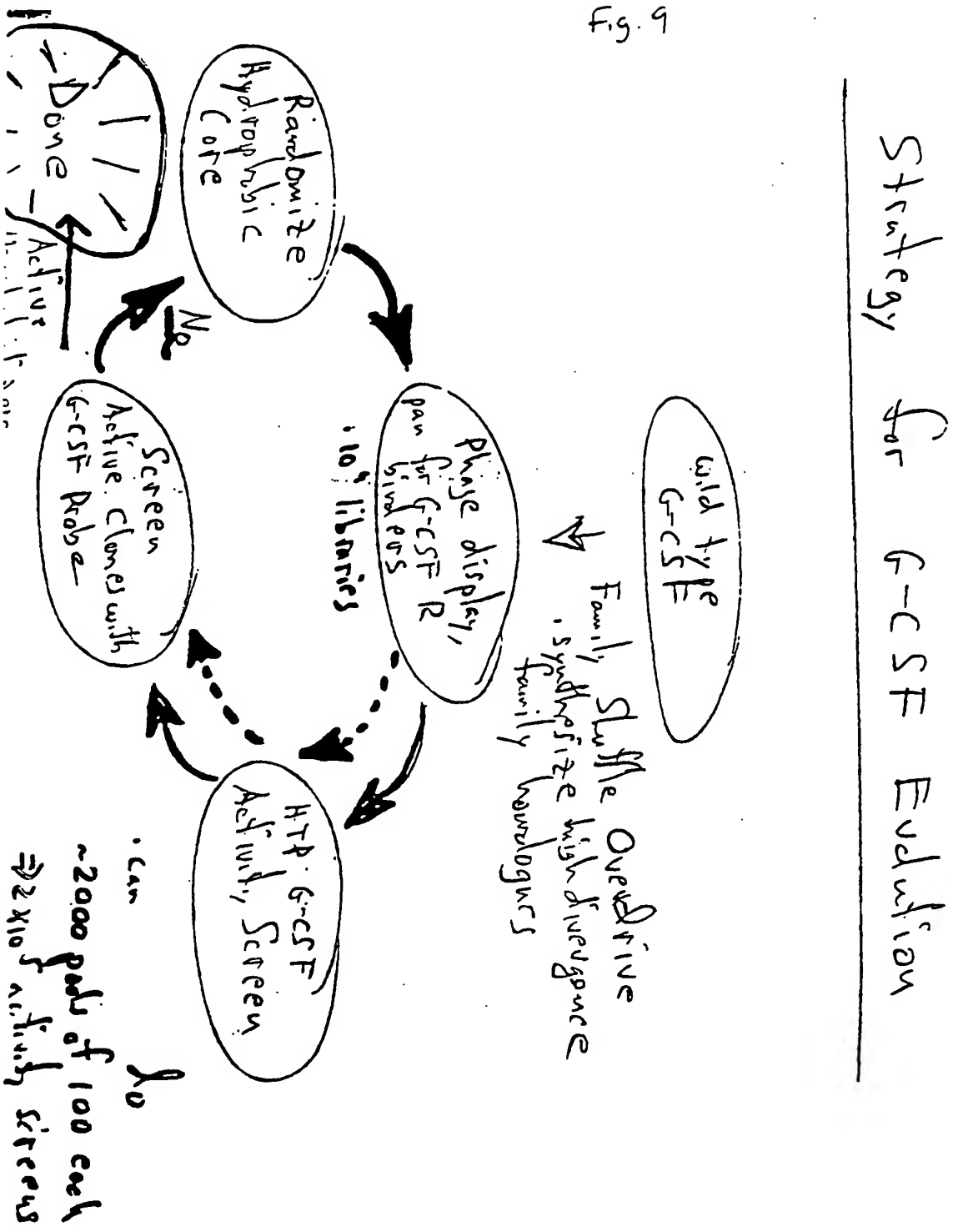


Fig. 10

Assembling of Alkaline Phosphatase

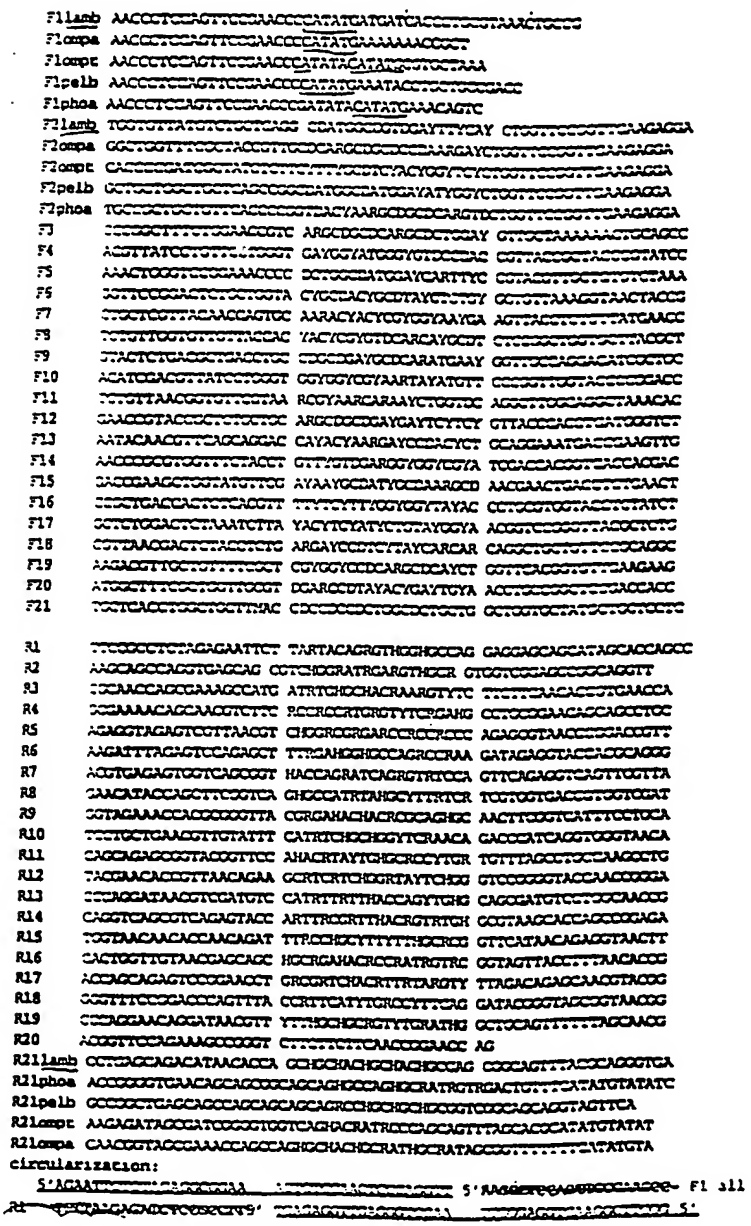
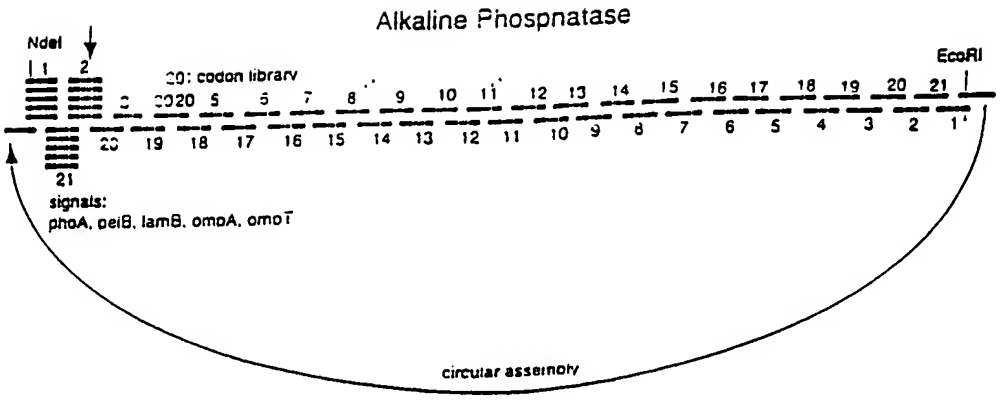


Fig. 11


Alkaline Phospnatase
NdeI
EcoRI
20: codon library
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
21
signals:
phoA, pelB, lamB, ompA, ompT
circular assembly

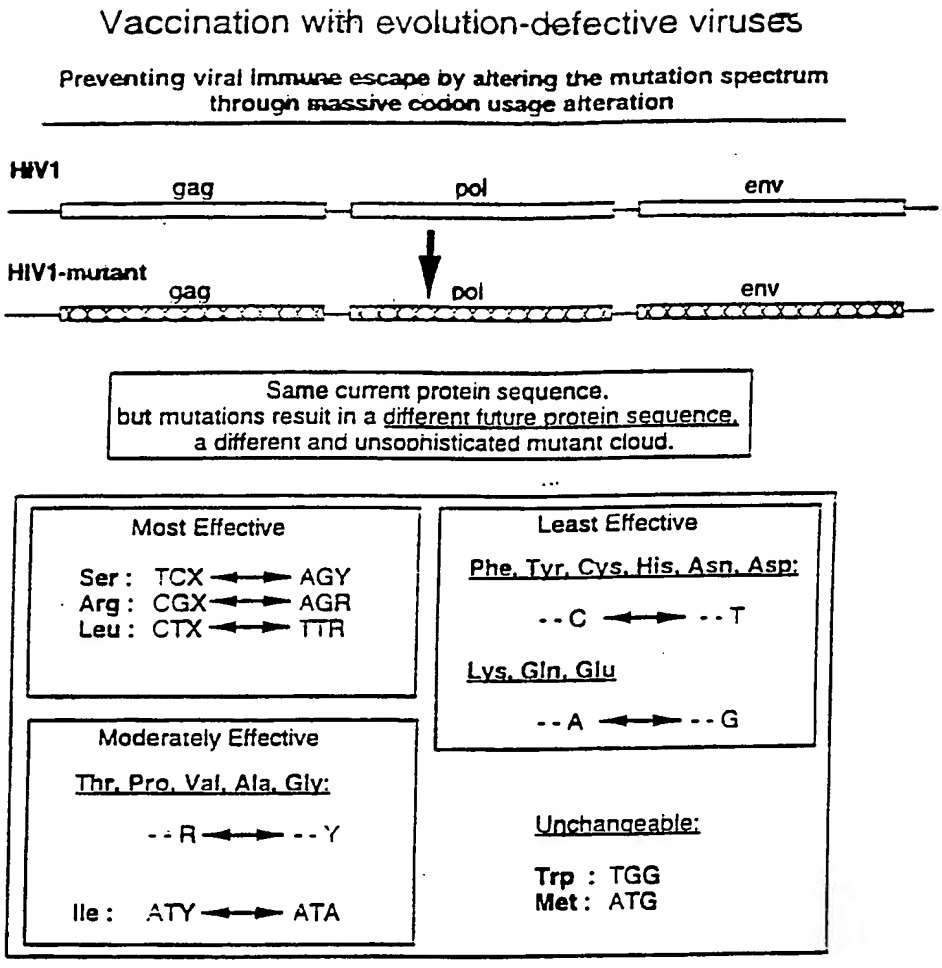
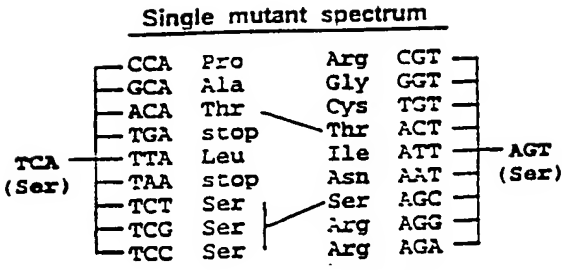

Vaccination with evolution-defective viruses
Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration
HIV1
gag
pol
env
HIV1-mutant
gag
pol
env
Same current protein sequence.
but mutations result in a different future protein sequence,
a different and unsophisticated mutant cloud.
Most Effective
Ser : TCX ◄► AGY
Arg : CGX ◄► AGR
Leu : CTX ◄► TTR
Least Effective
Phe, Tyr, Cys, His, Asn, Asp:
--C ◄► --T
Lys, Gln, Glu
--A ◄► --G
Moderately Effective
Thr, Pro, Val, Ala, Gly:
--R ◄► --Y
Ile : ATY ◄► ATA
Unchangeable:
Trp : TGG
Met : ATG

Fig. 12

Ser	TCX	—AGY
Arg	CGX	—AGR
Leu	CTX	—TTR



Common: 11% Ser, 11% Thr

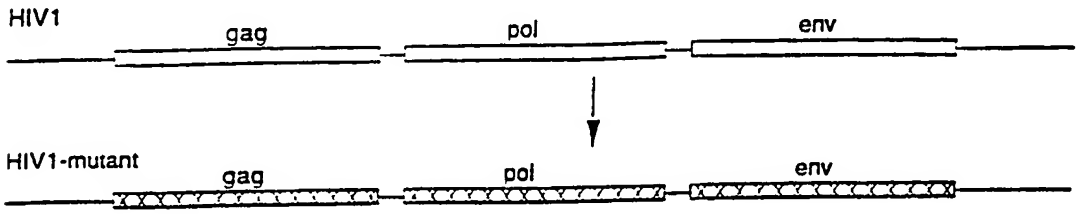
78% of mutations result in different amino acids

Fig. 13

Figure 14

Vaccination with evolution-defective viruses

Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration



Same current protein sequence,
but mutations result in a different future protein sequence,
a different and unsophisticated mutant cloud.

Maximizing the difference requires the following changes:

Most Effective

Ser : TCX ⟷ AGY
Arg : CGX ⟷ AGR
Leu : CTX ⟷ TTR

Moderately Effective

Thr, Pro, Val, Ala, Gly:

--R ⟷ --Y

Ile : ATY ⟷ ATA

Least Effective

Phe, Tyr, Cys, His, Asn, Asp:

--C ⟷ --T

Lys, Gln, Glu

--A ⟷ --G

Unchangeable:

Trp : TGG
Met : ATG

Fig. 15

Vaccination with evolution-defective viruses

Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration

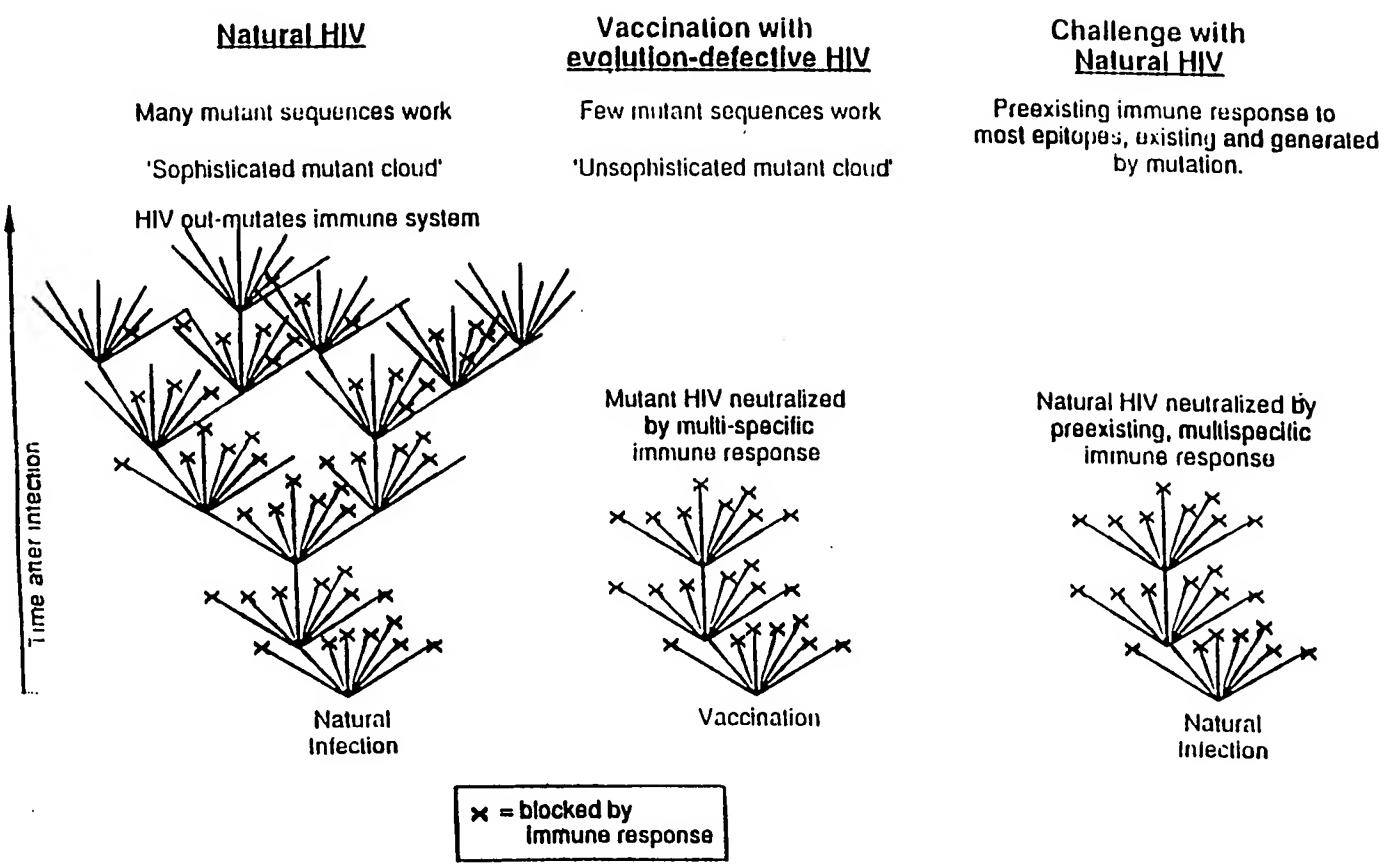


Fig. 16 A

Arg

AGA	CGT	AGA	CGC	AGA	CGA	AGA	CGG
TGA .	AGT S	TGA .	AGC S	TGA .	AGA S	TGA .	AGG S
CGA R	GGT G	CGA R	GGC G	CGA R	GGA G	CGA R	GGG G
GGA G	TGT C	GGA G	TGC C	GGA G	TGA .	GGA G	TGG W
AAA K	CAT H	AAA K	CAC H	AAA K	CAA Q	AAA K	CAG Q
ACA T	CCT P	ACA T	CCC P	ACA T	CCA P	ACA T	CCG P
ATA I	CTT L	ATA I	CTC L	ATA I	CTA L	ATA I	CTG L
AGC S	CGA R	AGC S	CGA R	AGC S	CGC R	AGC S	CGA R
AGT S	CGC R	AGT S	CGT R	AGT S	CGT R	AGT S	CGT R
AGG R	CGG R	AGG R	CGG R	AGG R	CGG R	AGG R	CGC R

AGG	CGT	AGG	CGC	AGG	CGA	AGG	CGG
TGG W	AGT S	TGG W	AGC S	TGG W	AGA S	TGG W	AGG S
CGG R	GGT G	CGG R	GGC G	CGG R	GGA G	CGG R	GGG G
GGG G	TGT C	GGG G	TGC C	GGG G	TGA .	GGG G	TGG W
AAG K	CAT H	AAG K	CAC H	AAG K	CAA Q	AAG K	CAG Q
ACG T	CCT P	ACG T	CCC P	ACG T	CCA P	ACG T	CCG P
ATG M	CTT L	ATG M	CTC L	ATG M	CTA L	ATG M	CTG L
AGA R	CGA R	AGA R	CGA R	AGA R	CGC R	AGA R	CGA R
AGC S	CGC R	AGC S	CGT R	AGC S	CGT R	AGC S	CGT R
AGT S	CGG R	AGT S	CGG R	AGT S	CGG R	AGT S	CGC R

Fig. 16B

Ser

AGT		TCT	
CGT	R	ACT	T
GGT	G	CCT	P
TGT	C	GCT	A
AAT	N	TAT	Y
ACT	T	TGT	C
ATT	I	TTT	F
AGA	R	TCA	S
AGG	R	TCG	S
AGC	S	TCC	S

AGT		TCC	
CGT	R	ACC	T
GGT	G	CCC	P
TGT	C	GCC	A
AAT	N	TAC	Y
ACT	T	TGC	C
ATT	I	TTC	F
AGA	R	TCA	S
AGG	R	TCG	S
AGC	S	TCT	S

AGT		TCA	
CGT	R	ACA	T
GGT	G	CCA	P
TGT	C	GCA	A
AAT	N	TAA	.
ACT	T	TGA	.
ATT	I	TTA	L
AGA	R	TCC	S
AGG	R	TCG	S
AGC	S	TCT	S

AGT		TCG	
CGT	R	ACG	T
GGT	G	CCG	P
TGT	C	GCG	A
AAT	N	TAG	.
ACT	T	TGG	W
ATT	I	TTG	L
AGA	R	TCA	S
AGG	R	TCC	S
AGC	S	TCT	S

AGC		TCT	
CGC	R	ACT	T
GGC	G	CCT	P
TGC	C	GCT	A
AAC	N	TAT	Y
ACC	T	TGT	C
ATC	I	TTT	F
AGA	R	TCA	S
AGG	R	TCG	S
AGT	S	TCC	S

AGC		TCC	
CGC	R	ACC	T
GGC	G	CCC	P
TGC	C	GCC	A
AAC	N	TAC	Y
ACC	T	TGC	C
ATC	I	TTC	F
AGA	R	TCA	S
AGG	R	TCG	S
AGT	S	TCT	S

AGC		TCA	
CGC	R	ACA	T
GGC	G	CCA	P
TGC	C	GCA	A
AAC	N	TAA	.
ACC	T	TGA	.
ATC	I	TTA	L
AGA	R	TCC	S
AGG	R	TCG	S
AGT	S	TCT	S

AGC		TCG	
CGC	R	ACG	T
GGC	G	CCG	P
TGC	C	GCG	A
AAC	N	TAG	.
ACC	T	TGG	W
ATC	I	TTG	L
AGA	R	TCA	S
AGG	R	TCC	S
AGT	S	TCT	S

Fig. 16

Leu

TTA		CTT	
ATA	I	ATT	I
CTA	L	GTT	V
GTA	V	TTT	F
TAA	.	CAT	H
TCA	S	CCT	P
TGA	.	CGT	R
TTG	L	CTA	L
TTC	F	CTC	L
TTT	F	CTG	L

TTA		CTC	
ATA	I	ATC	I
CTA	L	GTC	V
GTA	V	TTC	F
TAA	.	CAC	H
TCA	S	CCC	P
TGA	.	CGC	R
TTG	L	CTT	L
TTC	F	CTA	L
TTT	F	CTG	L

TTA		CTA	
ATA	I	ATA	I
CTA	L	GTA	V
GTA	V	TTA	F
TAA	.	CAA	H
TCA	S	CCA	P
TGA	.	CGA	R
TTG	L	CTT	L
TTC	F	CTC	L
TTT	F	CTG	L

TTA		CTG	
ATA	I	ATG	M
CTA	L	GTG	V
GTA	V	TTG	L
TAA	.	CAG	Q
TCA	S	CCG	P
TGA	.	CGG	R
TTG	L	CTT	L
TTC	F	CTC	L
TTT	F	CTA	L

TTG		CTT	
ATG	M	ATT	I
CTG	L	GTT	V
GTG	V	TTT	F
TAG	.	CAT	H
TCG	S	CCT	P
TGG	W	CGT	R
TTA	I	CTA	L
TTC	F	CTC	L
TTT	F	CTG	L

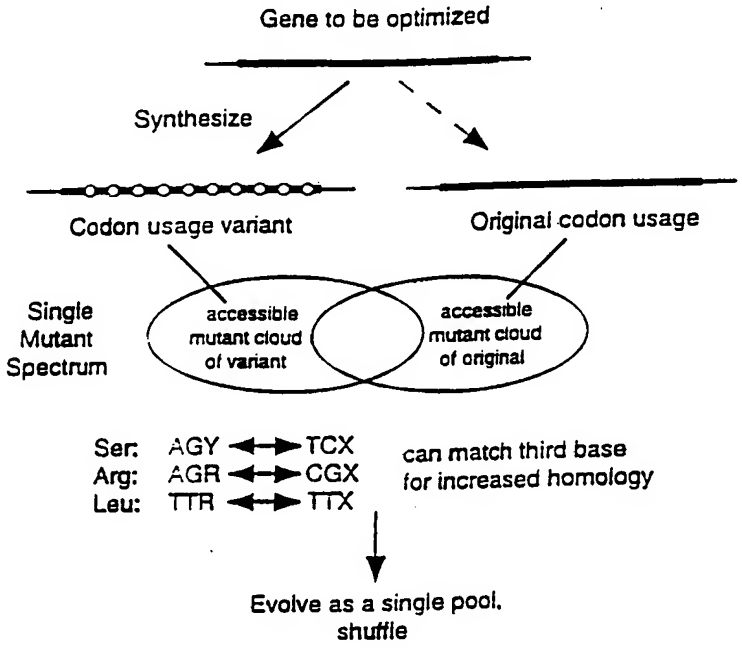
TTG		CTC	
ATG	M	ATC	I
CTG	L	GTC	V
GTG	V	TTC	F
TAG	.	CAC	H
TCG	S	CCC	P
TGG	W	CGC	R
TTA	I	CTT	L
TTC	F	CTA	L
TTT	F	CTG	L

TTG		CTA	
ATG	M	ATA	I
CTG	L	GTA	V
GTG	V	TTA	F
TAG	.	CAA	H
TCG	S	CCA	P
TGG	W	CGA	R
TTA	I	CTT	L
TTC	F	CTC	L
TTT	F	CTG	L

TTG		CTG	
ATG	M	ATG	M
CTG	L	GTG	V
GTG	V	TTG	L
TAG	.	CAG	Q
TCG	S	CCG	P
TGG	W	CGG	R
TTA	I	CTT	L
TTC	F	CTC	L
TTT	F	CTA	L

Fig. 17

Protein Evolution with Expanded Mutation Spectra



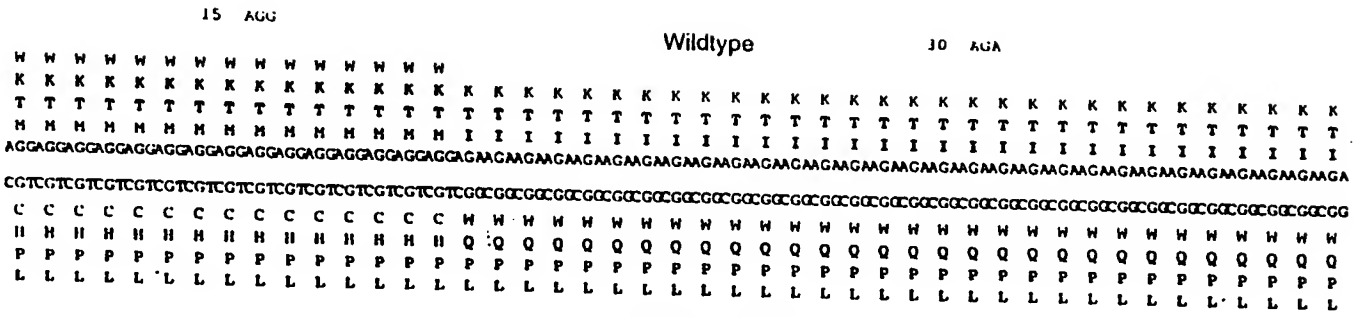
Best synthesis method: all possible codons based on codon synthesis

Fig 18A

Altering the evolution potential of sequences by codon modification

HIV envelope protein (857 amino acids)

Potential mutant resulting from wildtype sequence versus codon modified sequence in arg, leu and serine codons:



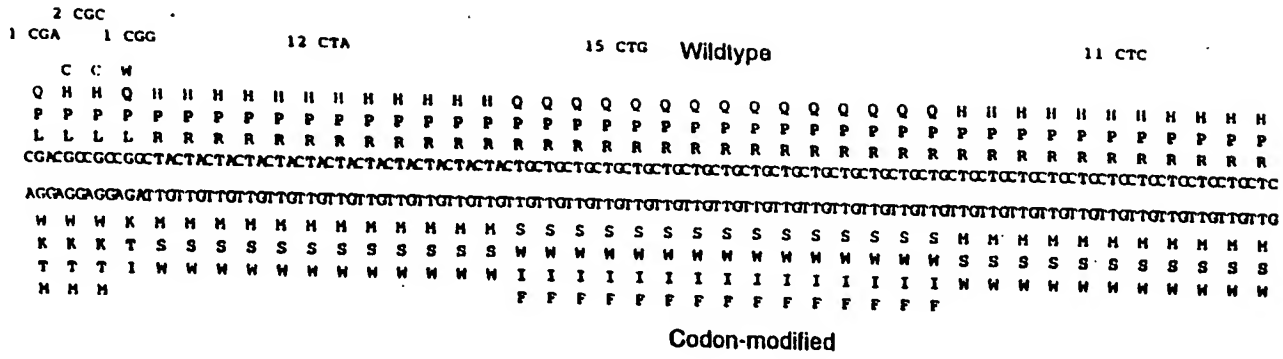


Fig. 18A

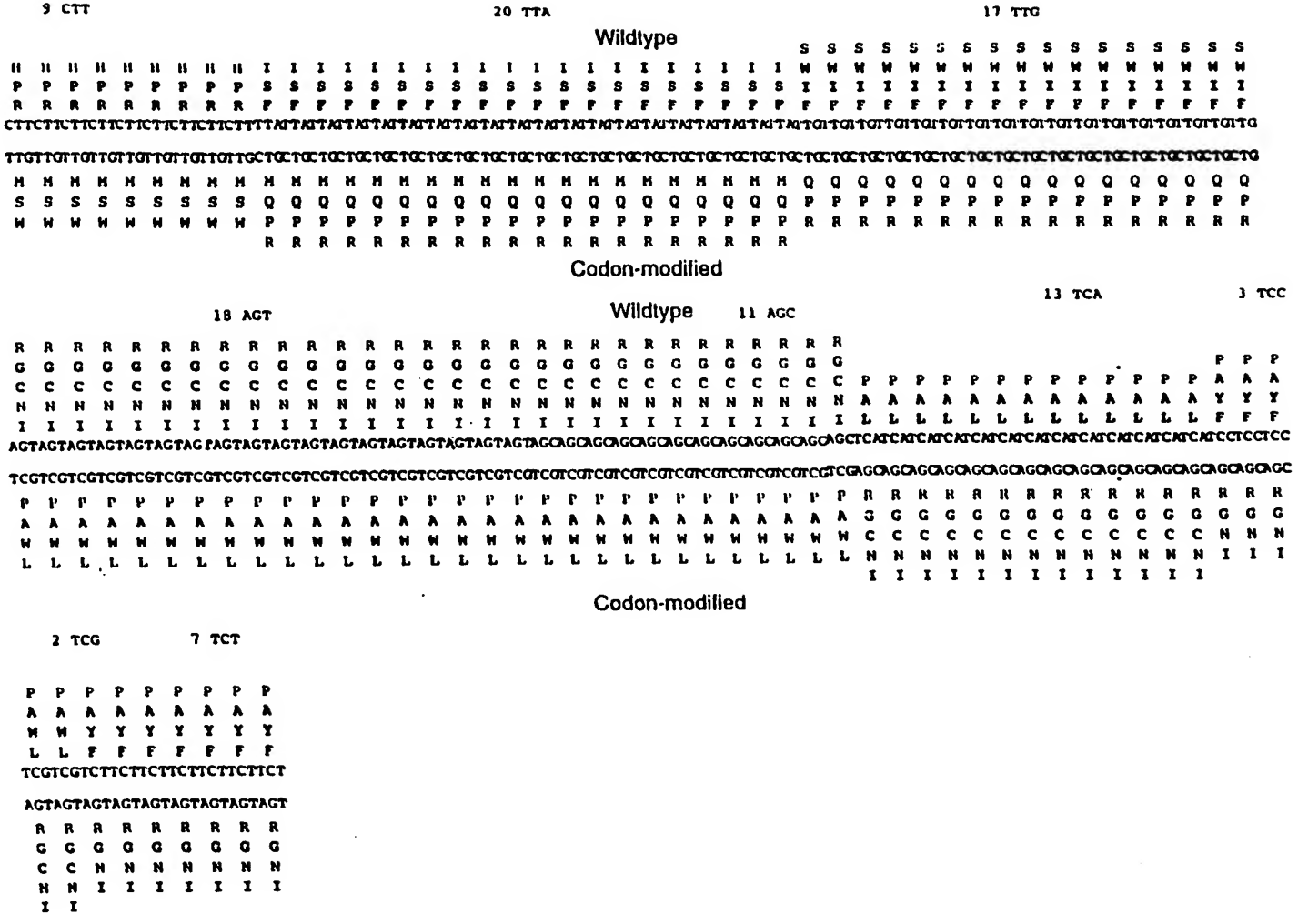


Fig 18B FINAL ALIGNMENT

NEW ENV VS OLD ENV

Saturday, August 17, 1 ō 2:03 PM
Lipman-Pearson Protein Alignment -
kTuple: 2; Gap Penalty: 4; Gap Length Penalty: 12

Seq1(1>509) new env pro	Seq2(1>509) old env protein	Similarity Index	Gap Number	Gap Length	Consensus Length
(1>509)	(1>509)	100.0	0	0	509

```
          ⌐10        ⌐20        ⌐30        ⌐40        ⌐50        ⌐60        ⌐70
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
          ¬10        ¬20        ¬30        ¬40        ¬50        ¬60        ¬70
          ⌐80        ⌐90        ⌐100       ⌐110       ⌐120       ⌐130       ⌐140
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
          ¬80        ¬90        ¬100       ¬110       ¬120       ¬130       ¬140
          ⌐150       ⌐160       ⌐170       ⌐180       ⌐190       ⌐200       ⌐210
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
          ¬150       ¬160       ¬170       ¬180       ¬190       ¬200       ¬210
          ⌐220       ⌐230       ⌐240       ⌐250       ⌐260       ⌐270       ⌐280
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
          ¬220       ¬230       ¬240       ¬250       ¬260       ¬270       ¬280
          ⌐290       ⌐300       ⌐310       ⌐320       ⌐330       ⌐340       ⌐350
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
          ¬290       ¬300       ¬310       ¬320       ¬330       ¬340       ¬350
          ⌐360       ⌐370       ⌐380       ⌐390       ⌐400       ⌐410       ⌐420
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
          ¬360       ¬370       ¬380       ¬390       ¬400       ¬410       ¬420
          ⌐430       ⌐440       ⌐450       ⌐460       ⌐470       ⌐480       ⌐490
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
          ¬430       ¬440       ¬450       ¬460       ¬470       ¬480       ¬490
          ⌐500
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
          ¬500
```

Fig 18C

DNA HOMOLOGY

Saturday, August 17, 1999 4:39 PM
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1>1527) new env	Seq2(1>1527) old env	Similarity Index	Gap Number	Gap Length	Consensus Length
(10>1454)	(10>1454)	61.7	0	0	1445

```
 10        20        30        40        50        60
GTBACBGTBTTTTACGGAGTTCCTGCTTGGAAAAACGCHACBATMCCDTTRTTTTGCGCT
GT AC GT TT TA GG GT CC GC TGGAA AA GC AC AT CC  T TT TG GC
GTAACAGTATTCTATGGTGTACCAGCATGGAAGAATGCGACAATTCCCCTCTTCTGTGCA
 10        20        30        40        50        60
 70        80        90        100       110       120
ACAAAAAACCGHGATACVTGGGGBACBACBCAGTGTCTCCCTGACAACGACGACTATAGY
AC AA AA  G GA AC TGGGG AC AC CA TG  T CC GA AA GA GA TA
ACCAAGAATAGGGACACTTGGGGAACAACACAATGCTTGCCAGATAATGATGATTACTCA
 70        80        90        100       110       120
 130       140       150       160       170       180
GAGCTHGCBATAAACGTAACTGAAGCATTCGACGCVTGGGACAACACBGTDACTGAGCAG
GA  T GC AT AA GT AC GA GC TT GA GC TGGGA AA AC GT AC GA CA
GAATTGGCAATCAATGTCACAGAGGCTTTTGATGCTTGGGATAATACAGTCACAGAACAA
 130       140       150       160       170       180
 190       200       210       220       230       240
GCTATTGAAGACGTHTGGAATTTRTTCGAGACTAGTATAAAACCATGCGTTAAGTTRACD
GC AT GA GA GT TGGAA  T TT GA AC    AT AA CC TG GT AA  T AC
GCAATAGAGGATGTGTGGAACCTCTTTGAAACATCCATTAAGCCCTGTGTAAAACTCACC
 190       200       210       220       230       240
 250       260       270       280       290       300
CCBTTGTGCATTGCTATGAGGTGCAACAAGACVGAAACBGACCGHTGGGGVCTCACTGGT
CC  T TG AT GC ATGAG TG AA AA AC GA AC GA  G TGGGG  T AC GG
CCACTATGTATAGCAATGAGATGTAATAAAACTGAGACAGATAGGTGGGGTTTGACAGGA
 250       260       270       280       290       300
 310       320       330       340       350       360
AATGCTGGTACBACBACBACBGCBATYACBACTACCGCTACBCCTTCVGTTGCBGAGAAC
AA GC GG AC AC AC AC GC AT AC AC AC GC AC CC    GT GC GA AA
AACGCAGGGACAACAACAACAGCAATAACAACAACAGCAACACCAAGTGTAGCAGAAAAT
 310       320       330       340       350       360
 370       380       390       400       410       420
GTYATYAACGAGTCAAACCCATGTATTAAGAACAACTCVTGCGCBGGDCTYGAGCAAGAA
GT AT AA GA    AA CC TG AT AA AA AA    TG GC GG  T GA CA GA
GTTATAAATGAAAGTAATCCTTGCATAAAAAATAATAGTTGTGCAGGCTTGGAACAGGAG
 370       380       390       400       410       420
 430       440       450       460       470       480
CCDATGATTGGVTGCAAGTTCAATATGACBGGCCTTAACCGCGATAAGAAAAAGGAGTAC
CC ATGAT GG TG AA TT AA ATGAC GG  T AA  G GA AA AA AA GA TA
CCCATGATAGGTTGTAAATTTAACATGACAGGGTTAAATAGGGACAAAAAGAAAGAATAT
 430       440       450       460       470       480
 490       500       510       520       530       540
AACGAGACBTGGTACAGTCGTGACCTTATATGCGATCAAAGYGCHAACGAGTCAGAATCC
AA GA AC TGGTA     G GA  T AT TG GA CA    GC AA GA    GA
AATGAAACATGGTATTCAAGAGATTTAATCTGTGAGCAGTCAGCGAATGAAAGTGAGAGT
 490       500       510       520       530       540
 550       560       570       580       590       600
AAGTGCTATATGCACCACTGCAATACDTCVGTAATACAGGAGAGTTGCGATAAACACTAC
AA TG TA ATGCA CA TG AA AC    GT AT CA GA    TG GA AA CA TA
AAATGTTACATGCATCATTGTAACACCAGTGTTATTCAAGAATCCTGTGACAAGCATTAT
 550       560       570       580       590       600
```

Fig. 18C

Saturday, August 17, 19[redacted] 4:39 PM

Page 2

Wilbur-Lipman DNA Alignment

Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1>1527) new env DNA	Seq2(1>1527) old env.pro	Similarity Index	Gap Number	Gap Length	Consensus Length
(10>1454)	(10>1454)	61.7	0	0	1445

```
 610       620       630       640       650       660
TGGGACGCVATACGTTTCCGCTATTGCGCTCCHCCBGGVTACGCVCTHTTRCGCTGCAAC
TGGGA GC AT  G TT  G TA TG GC CC CC GG TA GC  T  T  G TG AA
TGGGATGCTATTAGATTTAGATACTGTGCACCGCCAGGTTATGCTTTGCTTAGGTGTAAT
 610       620       630       640       650       660
 670       680       690       700       710       720
GACAGTAACTACCTBGGDTTCGCVCCVAATTGCAGYAAAGTTGTCGTAAGCAGYTGTACB
GA    AA TA  T GG TT GC CC AA TG    AA GT GT GT       TG AC
GATTCAAATTATTTAGGCTTTGCTCCTAACTGTTCTAAGGTAGTGGTTTCTTCATGCACA
 670       680       690       700       710       720
 730       740       750       760       770       780
CGBATGATGGAAACCCAGACGAGCACATGGTTCGGDTTTAACGGVACVCGHGCTGAGAAC
 G ATGATGGA AC CA AC    AC TGGTT GG TT AA GG AC  G GC GA AA
AGAATGATGGAGACGCAAACCTCTACTTGGTTTGGCTTCAATGGTACTAGGGCAGAAAAT
 730       740       750       760       770       780
 790       800       810       820       830       840
CGTACTTATATMTACTGGCACGGDAAGTCVAACCGTACGATTATATCGCTCAACAAATAC
 G AC TA AT TA TGGCA GG AA    AA  G AC AT AT     T AA AA TA
AGAACATACATTTATTGGCATGGCAAAAGTAATAGAACCATAATTAGCTTGAATAAGTAT
 790       800       810       820       830       840
 850       860       870       880       890       900
TACAACTTYACTATGCGTTGCCGTCGCCCTGGBAACAAAACBGTVCTBCCBGTAACGATA
TA AA  T AC ATG G TG  G  G CC GG AA AA AC GT  T CC GT AC AT
TATAATCTAACAATGAGATGTAGAAGACCAGGAAATAAGACAGTTTTACCAGTCACCATT
 850       860       870       880       890       900
 910       920       930       940       950       960
ATGACTGGCCTHGTDTTTCACAGYCAACCDATTAACGAACGTCCTAAGCAAGCDTGGTGT
ATG C GG  T GT TT CA    CA CC AT AA GA  G CC AA CA GC TGGTG
ATGTCAGGGTTGGTCTTCCATTCGCAGCCCATAAATGAGAGACCAAAACAGGCCTGGTGC
 910       920       930       940       950       960
 970       980       990       1000      1010      1020
TGGTTCGAGGGTTCGTGGAAGAAAGCGATACAAGAGGTHAAAGAGACDCTHGTGAAGCAC
TGGTT GA GG    TGGAA AA GC AT CA GA GT AA GA AC  T GT AA CA
TGGTTTGAAGGAAGCTGGAAAAAGGCCATCCAGGAAGTGAAGGAAACCTTGGTCAAACAT
 970       980       990       1000      1010      1020
 1030      1040      1050      1060      1070      1080
CCGCGCTACACHGGBACVAACGACACVCGHAAGATAAACTTGACTGCACCTGCBGGBGGB
CC  G TA AC GG AC AA GA AC  G AA AT AA  T AC GC CC GC GG GG
CCCAGGTATACGGGAACTAATGATACTAGGAAAATTAATCTAACAGCTCCAGCAGGAGGA
 1030      1040      1050      1060      1070      1080
 1090      1100      1110      1120      1130      1140
GACCCSGAGGTGACATTCATGTGGACTAACTGCAGRGGBGAGTTTCTBTACTGTAAGATG
GA CC GA GT AC TT ATGTGGAC AA TG  G GG GA TT  T TA TG AA ATG
GATCCAGAAGTCACTTTTATGTGGACAAATTGTCGAGGAGAATTCTTATATTGCAAAATG
 1090      1100      1110      1120      1130      1140
 1150      1160      1170      1180      1190      1200
AACTGGTTCTTRAACTGGGTBGAAGATCGBGATCAGAAAGGAGGGCGTTGGAAGCAGCAG
AA TGGTT  - AA TGGGT GA GA  G GA CA AA GG GG  G TGGAA CA CA
AATTGGTTTCTTAATTGGGTAGAGGACAGAGACCAAAAGGGTGGCAGATGCAAACAACAA
 1150      1160      1170      1180      1190      1200
```

Fig 18C

Saturday, August 17, 1[illegible] 4:39 PM
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1>1527)	Seq2(1>1527)	Similarity Index	Gap Number	Gap Length	Consensus Length
new env DNA	old env.pro				
(10>1454)	(10>1454)	61.7	0	0	1445

```
f1210       f1220       f1230       f1240       f1250       f1260
AACCGHAAGGAACAGCAAAAAAAGAACTACGTHCCBTGCCACATMCGBCAGATTATAAAT
AA  G AA GA CA CA AA AA AA TA GT CC TG CA AT  G CA AT AT AA
AATAGGAAAGAGCAACAGAAGAAAAATTATGTGCCATGTCATATTAGACAAATAATCAAC
t1210       t1220       t1230       t1240       t1250       t1260
f1270       f1280       f1290       f1300       f1310       f1320
ACCTGGCATAAGGTBGGDAAGAACGTBTACCTCCCACCCCGCGAGGGTGATTTRACBTGT
AC TGGCA AA GT GG AA AA GT TA  T CC CC  G GA GG GA  T AC TG
ACGTGGCACAAAGTAGGCAAAAATGTATATTTGCCTCCTAGGGAAGGAGACCTGACATGC
t1270       t1280       t1290       t1300       t1310       t1320
f1330       f1340       f1350       f1360       f1370       f1380
AACAGYACVGTTACATCATTGATTGCTGAAATYGACTGGATWAACTCDAACGAAACGAAC
AA    AC GT AC     T AT GC GA AT GA TGGAT AA    AA GA AC AA
AATTCCACTGTAACTAGTCTCATAGCAGAGATAGATTGGATCAATAGCAATGAGACCAAT
t1330       t1340       t1350       t1360       t1370       t1380
f1390       f1400       f1410       f1420       f1430       f1440
ATAACGATGTCVGCBGAAGTHGCBGAGTTYTACAGGCTCGAACTCGGTGACTATAAGCTB
AT AC ATG   GC GA GT GC GA  T TA  G  T GA  T GG GA TA AA  T
ATCACCATGAGTGCAGAGGTGGCAGAACTGTATCGATTGGAGTTGGGAGATTACAAATTA
t1390       t1400       t1410       t1420       t1430       t1440
f1450
ATYGA
AT GA
ATAGA
t1450
```

Saturday, August 17, 1996 2 :3 PM Fig 18d Page 1
1.6 kb Env mutagenized
Created: Friday, August 16, 1996 9:39 AM

MUTAGENIZED GENE IN CONTEXT

```
        10        20        30        40
aactccgaagaagactaaggctaatccatttcctgcatca 40
aacaagtaagtatgggatgtcttgggaatcagctgcttat 80
cgcgctcttgctagtaagtgttttagagatttgttgtGTv 120
CAgTAcGTbACbGTbTTtTAcGGaGTtCCTGCtTGGAAaA 160
AcGChACbATmCCdttrTTtTGcGCtACaAAaAAccghGA 200
      210       220       230       240
tACvTGGGGbACbACbCAgTGtctcCCtGAcAAcGAcGAc 240
TAtagyGAgcthGCbATaAAcGTaACtGAaGCaTTcGAcG 280
CvTGGGAcAAcACbGTdACtGAgCAgGCtATtGAaGAcGT 320
hTGGAAtttrTTcGAgACtagtATaAAaCCaTGcGTtAAg 360
ttrACdCCbttgTGcATtGCtATGAGgTGcAAcAAgACvG 400
      410       420       430       440
AaACbGAccgnTGGGGvctcACtGGtAAtGCtGGtACbAC 440
bACbACbGCbATyACbACtACcGCtACbCCttcvGTtGCb 480
GAgAAcGTyATyAAcGAgtcaAAcCCaTGtATtAAgAAcA 520
ActcvTGcGCbGGdctyGAgCAaGAaCCdATGATtGGvTG 560
cAAgTTcAAtATGACbGGccttAAccgcGAtAAgAAaAAg 600
      610       620       630       640
GAgTAcAAcGAgACbTGGTAcagtcgtGAccttATaTGcG 640
AaCAaagyGChAAcGAgtcaGAatccAAgTGcTAtATGCA 680
cCAcTGcAAtACdtcvGTaATaCAgGAgagtTGcGAtAAa 720
CAcTAcTGGGAcGCvATacgtTTccgcTAtTGcGCtCChC 760
CbGGvTAcGCvcthttrcgcTGcAAcGAcagtAAcTAcct 800
      810       820       830       840
bGGdTTcGCvCCvAAtTGcagyAAaGTtGTcGTaagcagy 840
TGtACbcgbATGATGGAaACcCAgACgagcACaTGGTTcG 880
GdTTtAAcGGvACvcghGCtGAgAAccgtACtTAtATmTA 920
cTGGCAcGGdAAgtcvAAccgtACgATtATatcgctcAAc 960
AAaTAcTAcAActtgACtATGcgtTGccgtcgcCCtGGbA 1000
      1010      1020      1030      1040
AcAAaACbGTvctbCCbGTaACgATaATGagtGGccthGT 1040
dTTtCAcagyCAaCCdATtAAcGAacgtCCtAAgCAaGCd 1080
TGGTGtTGGTTcGAgGGttcgTGGAAgAAaGCgATaCAaG 1120
AgGThAAaGAgACdcthGTgAAgCAcCCgcgcTAcAChGG 1160
bACvAAcGAcACvcgnAAgATaAActtgACtGCaCCtGCb 1200
```

Figure 19

ENV Top oligos

1T CTAGTAAGTGTTTTAGAGATTTGTTGT

2T TTACGGAGTTCCTGCTTGGAAAAACGCHACBATMCCDTTRTTTTGCGCTACAAAAAACCG

3T AGTGTCTCCCTGACAACGACGACTATAGYGAGCTHGCBATAAACGTAACTGAAGCATTCG

4T ACTGAGCAGGCTATTGAAGACGTHTGGAATTTRTTCGAGACTAGTATAAAACCATGCGTT

5T TGCTATGAGGTGCAACAAGACVGAAACBGACCGHTGGGGVCTCACTGGTAATGCTGGTAC

6T CTACCGCTACTCCTTCAGTTGCBGAGAACGTVATYAACGAGTCAAACCCATGTATTAAGA

7T GAGCAAGAACCGATGATTGGVTGCAAGTTCAATATGACBGGCCTTAACCGCGATAAGAAA

8T GTACAGTCGTGACCTTATATGCGAACAAAGYGCHAACGAGTCAGAATCCAAGTGCTATAT

9T TAATACAGGAGAGTTGCGATAAACACTACTGGGACGCVATACGTTTCCGCTATTGCGCTC

10T CGCTGCAACGACAGTAACTACCTBGGDTTCGCVCCVAATTGCAGTAAAGTTGTCGTAAGC

11T AACCCAGACGAGCACATGGTTCGGDTTTAACGGVACVCGHGCTGAGAACCGTACTTATAT

12T ACCGTACGATTATATCGCTCAACAAATACTACAACTTGACTATGCGTTCCGGTCGCCCTG

13T GTAACGATAATGAGTGGCCTHGTDTTTCACAGYCAACCDATTAACGAACGTCCTAAGCAA

14T TTCGTGGAAGAAAGCGATACAAGAGGTHAAAGAGACDCTHGTGAAGCACCCCGCCTACAC

15T AGATAAACTTGACTGCAACCTGCBGGBGGBGACCCBGAGGTGACATTCATGTGGACTAACT

16T TGTAAGATGAACTGGTTCTTRAACTGGGTBGAAGATCGBGATCAGAAAGGAGGGCGTTGG

17T ACAGCAAAAAAAGAACTACGTHCCBTGCCACATMCGBCAGATTATAAATACCTGGCATAA

18T TCCCACCCCGCGAGGGTGATTTRACBTGTAACAGYACVGTTACATCATTGATTGCTGAAA

19T GAAACGAACATAACGATGTCVGCBGAAGTHCCBGAGTTRTACAGGCTCGAACTCGGTGAC

20T ACCTATAGGGCTCGCGCCGACBTCVGTBCGBCGHTATACDACTACAGGAGCGAGTCGTAA

Fig 9 (cont.)

ENV BOTTOM OLIGOS

1B CCCCVCGTTTGTTACGACTCGCTCCTGTAGT

2B CTGGCGCGAGCCCTATAGGTGTKATTTCRATVAGCTTATAGTCACCGAGTTCGAGCCTGT

3B GACATCGTTATGTTCGTTTCGTTBGAGTTWATCCAGTCRATTTCAGCAATCAATGATGTA

4B ATCACCCTCGCGGGGTGGGAGGTAVACGTTCTTBCCVACCTTATGCCAGGTATTTATAAT

5B CGTAGTTCTTTTTTTGCTGTTCCTTDGGTTCTGCTGCTTCCAACGCCCTCCTTTCTGAT

6B AAGAACCAGTTCATCTTACAGTAVAGAAACTCVCCYCTGCAGTTAGTCCACATGAATGTC

7B AGGTGCAGTCAAGTTTATCTTDCGBGTGTCGTTBGTVCCDGTGTAGCGCGGGTGCTTCAC

8B GTATCGCTTTCTTCCACGAACCCTCGAACCAACACCAHGCTTGCTTAGGACGTTCGTTAA

9B AGGCCACTCATTATCGTTACVGGVAGBACVGTTTTGTTVCCAGGGCGACGGCAACGCATA

10B GAGCGATATAATCGTACGGTTBGACTTBCCGTGCCAGTAKATATAAGTACGGTTCTCAGC

11B ACCATGTGCTCGTCTGGGTTTCCATCATVCGVGTACARCTGCTTACGACAACTTTACTGC

12B TAGTTACTGTCGTTGCAGCGYAADAGBGCGTABCCVGGDGGAGCGCAATAGCGGAAACGT

13B ATCGCAACTCTCCTGTATTACBGAHGTATTGCAGTGGTGCATATAGCACTTGGATTCTGA

14B ATATAAGGTCACGACTGTACCAVGTCTCGTTGTACTCCTTTTTCTTATCGCGGTTAAGGC

15B CCAATCATCGGTTCTTGCTCRAGHCCVGCGCABGAGTTGTTCTTAATACATGGGTTTGAC

16B AACTGAAGGAGTAGCGGTAGTVGTRATVGCVGTVGTVGTVGTACCAGCATTACCAGTGAG

17B TCTTGTTGCACCTCATAGCAATGCACAAVGGHGTYAACTTAACGCATGGTTTTATACTAG

18B TCTTCAATAGCCTGCTCAGTHACVGTGTTGTCCCABCGCGTCGAATGCTTCAGTTACGTTT

19B GTCGTTGTCAGGGAGACACTGVGTVGTVCCCCABGTATCDCGGTTTTTTGTAGCGCAAAA

20B TCCAAGCAGGAACTCCGTAAAAVACVGTVACGTACTGBACACAACAAATCTCTAAAACAC

CORRECTED VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
C12N 15/10, 15/11, 5/10, 7/04, A61K 39/21, C07B 61/00, G06F 19/00, B01J 19/00 // C12N 15/12, 15/55, 15/49

A2

(11) International Publication Number: WO 00/18906

(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22588

(22) International Filing Date: 28 September 1999 (28.09.99)

(30) Priority Data:

60/102,362	29 September 1998 (29.09.98)	US
60/117,729	29 January 1999 (29.01.99)	US
60/118,813	5 February 1999 (05.02.99)	US
60/141,049	24 June 1999 (24.06.99)	US

(71) Applicant *(for all designated States except US)*: MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: PATTEN, Phillip, A. [US/US]; Apartment 506, 2680 Fayette Drive, Mountain View, CA 94040 (US). LIU, Lu [US/US]; 519 Skiff Circle, Redwood City, CA 94056 (US). STEMMER, Willem, P., C. [NL/US]; 108 Kathy Court, Los Gatos, CA 95030 (US).

(74) Agents: QUINE, Jonathan, Alan; Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

Without international search report and to be republished upon receipt of that report.

(54) Title: SHUFFLING OF CODON ALTERED GENES

(57) Abstract

Methods of recombining codon–altered libraries of nucleic acids are provided. The nucleic acids can include conservative or non–conservative modifications of coding sequences, in addition to codon alterations, as compared with wild–type sequences. In addition to making new proteins, methods of generation vectors with reduced rates of reversion to wild–type and attenuated viruses are also provided.

*(Referred to in PCT Gazette No. 14/2000, Section II)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

"SHUFFLING OF CODON ALTERED GENES," Attorney Docket No. 02-028500, by Patten and Stemmer, filed 09/29/98, and 60/117,729, "SHUFFLING OF CODON ALTERED GENES," Attorney Docket No. 02-028510, by Patten and Stemmer, filed January 29, 1999. The application is also related to USSN 60/118,813 "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296, filed February 5, 1999; and USSN 60/141,049 "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296-1, filed June 24, 1999.

BACKGROUND

The genetic code is highly degenerate. Every DNA/RNA triplet (codon) encoding an amino acid can typically be altered, with the exception of ATG/AUG (coding for methionine) and TGG/UGG (coding for Tryptophan), without altering the sequence of the protein encoded by the corresponding nucleic acid sequence. Roughly, on average (the distribution of amino acids varies from protein to protein), each coding triplet can be substituted about 3 different ways, since there are 61 codons encoding 20 amino acids (there are 3 additional triplets encoding stop codons, for a total of 64 codons encoding 20 amino acids). This represents a possible sequence diversity of approximately 3^n possible sequences which encode a given protein, where n is the length of the protein in amino acids. As can easily be seen, for proteins of even modest length, the number of possible nucleic acids which can encode the protein exceeds the number of physical particles in the universe (estimated at about 10^{80} particles).

This tremendous potential coding sequence space for individual proteins has interesting evolutionary implications. For example, hypermutable viruses such as HIVs and other retroviruses typically stay one step ahead of the host immune system by accumulating non-random mutations based, in part, upon the particular codons used to encode recognition molecules, e.g., in the envelope portion of the virus. The mutations are non-random because viruses are selected for the ability to mutate to forms which are not quickly recognized by the

using recursive shuffling methods. The nucleic acids or the encoded protein are then screened for a desirable property.

Thus, the invention provides methods of making codon altered nucleic acids. In the methods, a first nucleic acid sequence encoding a first polypeptide sequence is selected. A plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide, or a modified form thereof, are then selected (e.g., a library of codon altered nucleic acids can be selected in a biological assay which recognizes library components or activities), and the plurality of codon altered nucleic acid sequences is recombined to produce a target codon altered nucleic acid encoding a second protein. The target codon altered nucleic acid is then screened for a detectable functional or structural property, optionally including comparison to the properties of the first polypeptide. The goal of such screening is to identify a polypeptide that has a structural or functional property equivalent or superior to the first polypeptide. A nucleic acid encoding such a polypeptide can be used in essentially any procedure desired, including introducing the target codon altered nucleic acid into a cell, vector, virus, attenuated virus (e.g., as a component of a vaccine or immunogenic composition), transgenic organism, or the like.

Kits and compositions for practicing the methods are also provided, including one or more of: cell recombination mixtures and substrates (e.g., nucleic acids with altered codon usage), containers, instructional material for practicing the methods, or the like.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a nucleic acid/amino acid sequence of a part of the monkey EPO gene, which is similar to the human EPO gene.

Figure 2 shows an example of a codon altered EPO nucleic acid sequence.

Figure 3 shows an alignment of naturally occurring EPOs.

Figure 4 is a schematic of the human EPO wobble sequence space.

Figure 5 is a schematic of Mammalian EPO Family-Wobble Sequence Space.

Figure 6 is a sequence alignment of G-CSF homologs, with species information.

Figure 7 is a sequence alignment of G-CSF homologs, with differences broken out.

Figure 8 is a sequence alignment showing the hydrophobic core residues of human G-CSF (blacked out).

Figure 9 is a schematic showing the shuffling strategy for G-CSF.

Figure 10 is a list of oligos used to make a codon altered alkaline phosphatase.

Figure 11 is a map of oligos used to make a codon altered alkaline phosphatase.

Figure 12 is a schematic of vaccination with evolution defective viruses.

Figure 13 is a schematic of different mutations that result from different codon types for ser, arg, and leu.

Figure 14 is a schematic of vaccination with evolution defective viruses.

Figure 15 is a schematic of vaccination with evolution defective viruses showing sophisticated versus non-sophisticated "mutant clouds."

Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Fig. 17 is a schematic of protein evolution with expanded mutation spectra.

Fig. 18, panels A-D show codon altered forms of Env.

Fig. 19 is a list of oligos in one application for synthesis of HIV Env.

DEFINITIONS

Unless clearly indicated to the contrary, the following definitions supplement definitions of terms known in the art.

As used herein, a "recombinant" nucleic acid is a nucleic acid produced by recombination between two or more nucleic acids, or any nucleic acid made by an *in vitro* or artificial process. The term "recombinant" when used with reference to a cell indicates that the cell comprises (and optionally replicates) a heterologous nucleic acid, or expresses a peptide or protein encoded by a heterologous nucleic acid. Recombinant cells can contain genes that are not found within the native (non-recombinant) wild-type form of the cell. Recombinant cells can also contain genes found in the native form of the cell where the genes are modified and re-introduced into the cell by artificial means. The term also encompasses cells that contain a nucleic acid endogenous to the cell that has been artificially modified without removing the nucleic acid from the cell; such modifications include those obtained by gene replacement, site-specific mutation, chimeraplasty, and related techniques.

A "codon altered" nucleic acid is a first nucleic acid that encodes a first polypeptide similar or identical to a naturally occurring polypeptide encoded by a naturally occurring nucleic acid, where the first nucleic acid utilizes a plurality of codons to encode the first polypeptide, which differ from the codons of the naturally occurring nucleic acid that encode the naturally occurring polypeptide.

A "nucleic acid sequence" refers to either a nucleic acid (e.g., RNA, DNA or modified form thereof, in isolated, recombinant or native form) or to a representation of the nucleic acid such as a sequence of letters indicating the primary structure (sequence) of the nucleic acid.

A "polypeptide sequence" refers to either a polypeptide (or modified form thereof, in isolated, recombinant or native form) or to a representation of the polypeptide such as a sequence of letters or other character string information indicating the primary structure (amino acid sequence) of the polypeptide.

A "modified form" of a reference polypeptide is a target polypeptide which has a similar, but not identical, sequence to the reference polypeptide. The sequence of the target polypeptide can differ from the reference polypeptide by conservative or non-conservative substitutions of the reference polypeptide sequence. As noted in more detail, *supra*, different nucleic acids encoding different target polypeptides having different non-conservative substitutions relative to the reference polypeptide can be recombined to produce a recombined nucleic acid encoding a target polypeptide more similar to the reference polypeptide.

A "plurality of forms" of a selected nucleic acid refers to a plurality of homologs of the nucleic acid. The homologs can be from naturally occurring homologs (e.g., two or more homologous genes, or derivatives thereof) or by artificial synthesis of one or more nucleic acids having related sequences, or by modification of one or more nucleic acid to produce related nucleic acids. Nucleic acids are homologous when they are derived, naturally or artificially, from a common ancestor sequence. During natural evolution, this occurs when two or more descendent sequences diverge from a parent sequence over time, i.e., due to mutation and natural selection. Under artificial conditions, divergence occurs, e.g., in one of two ways. First, a given sequence can be artificially recombined with another sequence, as occurs, e.g., during typical cloning, to produce a descendent nucleic acid.

Alternatively, a nucleic acid can be synthesized *de novo*, by synthesizing a nucleic acid which varies in sequence from a given parental nucleic acid sequence.

When there is no explicit knowledge about the ancestry of two nucleic acids, homology is typically inferred by sequence comparison between two sequences. Where two nucleic acid sequences show sequence similarity it is inferred that the two nucleic acids share a common ancestor. The precise level of sequence similarity required to establish homology varies in the art depending on a variety of factors. For purposes of this disclosure, two sequences are considered homologous where they share sufficient sequence identity to allow recombination to occur between two nucleic acid molecules, or when codon changes can be made which would result in two or more nucleic acids having the ability to recombine. Typically, nucleic acids require regions of close similarity spaced roughly the same distance apart to permit recombination to occur.

The terms "identical" or percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or polypeptides refers to two or more sequences or subsequences that have at least about 40%, 50%, 60%, or preferably about 70% or 80% or more, or most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Such "substantially identical" sequences are typically considered to be homologous. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer,

subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al., infra*).

One example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation

(E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

Another indication that two nucleic acid sequences are substantially identical/homologous is that the two molecules hybridize to each other under stringent conditions. The phrase "hybridizing specifically to," refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions, including when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences and sequences with higher G:C content remain hybridized at higher temperatures (or at lower salt). An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York.

Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but not to unrelated (non-homologous) sequences.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42 °C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15M NaCl at 72 °C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65 °C for 15 minutes (*see,* Sambrook, *infra.,* for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, *e.g.,* more than 100 nucleotides, is 1x SSC at 45 °C for 15 minutes. An example low stringency wash for a duplex of, *e.g.,* more than 100 nucleotides, is 4-6x SSC at 40 °C for 15 minutes. For short probes (*e.g.*, about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 40 °C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. If the signal to noise ratio is less than 2x binding of an unrelated probe (e.g., a nucleic acid encoding a non-homologous protein), the nucleic acids at issue do not hybridize under stringent conditions. Similarly, if the signal to noise ratio is less than 25% as high as that observed for a perfectly matched probe under stringent conditions, the nucleic acids do not "hybridize under stringent conditions" as that term is used herein. This does not apply to highly stringent conditions, as the stringency can theoretically be increased until only a perfectly matched probe will hybridize.

In one example hybridization procedure, a target nucleic acid to be probed is blotted onto a filter by any conventional method. An unrelated nucleic acid such as a plasmid

vector (assuming that the target nucleic acid has no homology with the target nucleic acid) is also blotted, in approximately equal amounts onto the filter. The filter is probed with a labeled probe complementary to the target nucleic acid. The experiment is repeated at gradually increasing stringency of hybridization and wash conditions until signal from the hybridization of the labeled probe to the complementary target is 10-100X as high as to the unrelated plasmid vector nucleic acid. Once these conditions are determined as described above, a test nucleic acid is probed under the same conditions as the target. If signal from the labeled probe is 25% as high or higher than the signal from binding of the probe to the target, the test nucleic acid "hybridizes under stringent conditions" to the probe. If the signal is less than 25% as high, the test nucleic acid does not hybridize under stringent conditions to the probe.

Nucleic acids which do not hybridize to each other under stringent conditions are still recognizable as variant forms of a nucleic acid when the polypeptides they encode are substantially identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. Such nucleic acids are not functionally equivalent, as described in detail herein, due to differences in mRNA folding, alterations of regulatory sequences and the like.

Another indication that two nucleic acid sequences or polypeptides are variant forms is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid, as tested by polyclonal antisera generated to the first polypeptide. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

"Conservatively modified variations" of a particular polynucleotide sequence are those polynucleotide variations that encode identical or essentially identical amino acid sequences, or where the polynucleotide does not encode an amino acid sequence, which encode essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such

nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every polynucleotide sequence described herein which encodes a polypeptide also optionally describes every possible silent variation, except where otherwise noted. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the codon for methionine, and TGG, which is ordinarily the codon for tryptophan) can be modified to yield a peptide which is structurally identical.

Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following five groups each contain amino acids that are conservative substitutions for one another:

Aliphatic: Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I); Aromatic: Phenylalanine (F), Tyrosine (Y), Tryptophan (W); Sulfur-containing: Methionine (M), Cysteine (C);Basic: Arginine (R), Lysine (K), Histidine (H); Acidic: Aspartic acid (D), Glutamic acid (E), Asparagine (N), Glutamine (Q). *See also*, Creighton (1984) *Proteins*, W.H. Freeman and Company. In addition, individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservatively modified variations." Sequences that differ by conservative variations are generally homologous.

The term "isolated", when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular or other components (e.g., library components) with which it is associated in the natural state.

The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.* degenerate codon substitutions) and complementary sequences and as well as the sequence

explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.* (1991) *Nucleic Acid Res. 19*: 5081; Ohtsuka *et al.* (1985) *J. Biol. Chem. 260:* 2605-2608; Cassol *et al.* (1992) ; Rossolini *et al.* (1994) *Mol. Cell. Probes 8*: 91-98). The term nucleic acid is generic to the terms "gene", "DNA," "cDNA", "oligonucleotide," "RNA," "mRNA," and the like.

"Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the gene and detection of such derived products is indicative of the presence and/or abundance of the original gene and/or gene transcript in a sample.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence.

A "recombinant expression cassette" or simply an "expression cassette" is a nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gene in hosts compatible with such sequences. Expression cassettes include at least promoters and optionally, transcription termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (*e.g.*, a nucleic acid encoding a desired polypeptide), and a promoter. Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell. Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

DETAILED DISCUSSION OF THE INVENTION

In the present invention, the sequence diversity of substrates for DNA shuffling procedures is increased by using codon-altered nucleic acids as templates and/or by using templates that encode proteins with conservative or non-conservative amino acid modifications as compared to a selected wild-type protein.

These codon altered nucleic acids can be chemically synthesized (e.g., using standard artificial synthetic protocols, e.g., those typically used by commercial sources from which nucleic acids can be ordered), or can be made using any of a variety of methods herein of available to one of skill. For example, oligonucleotide fragments can be made which correspond to a codon altered nucleic acid which is desired using standard synthetic methods, followed by polymerase and/or ligase mediated oligonucleotide ligation/recombination protocols to generate full-length nucleic acids.

The combination of codon usage modifications and coding modifications can be extensive enough to reduce or, under stringent conditions, even eliminate the hybridization of the codon-altered nucleic acids to a nucleic acid which naturally encodes the selected protein. This dramatically alters the mutations which result from possible single nucleotide mutations, providing access to greater diversity for DNA shuffling protocols.

In addition, the recombination and selection of such nucleic acids during DNA shuffling procedures can result not only in access to a different set of possible mutations, but can also result in modified forms of transcriptional or translational regulation, alterations in nucleic acid localization, mRNA stability and the like. Furthermore, the modified hybridization properties of codon altered nucleic acids leads to alterations in the ability of the nucleic acids to hybridize with potential recombination partners, altering, and ultimately increasing, the available recombination diversity during shuffling.

Furthermore, "family shuffling" using codon-altered substrates even further increases the possible sequence diversity of the starting materials for recombination. As currently practiced, family shuffling methods involve shuffling nucleic acids encoding sequence variants of a given protein (e.g., species or allele homologs). In the present methods, this procedure is modified by generating codon-altered versions of the sequence variants to access additional molecular diversity during recombination. Additional diversity is achieved by conservatively and non-conservatively modifying the starting nucleic acids to encode non-naturally occurring sequence variants. Family shuffling can be performed even using homologs of relatively low identity. In such cases, codons may be changed in one or more of the family members to increase the level of identity between the members, thereby increasing their ability to recombine using the methods of this invention.

Gene shuffling and family shuffling provide two of the most powerful methods available for improving and "migrating" (gradually changing the type of reaction, substrate or activity of a selected protein such as an enzyme, or regulation or structure of an expressed component) the functions of proteins. In family shuffling, homologous sequences, e.g., from different species, chromosomal positions, or due to synthetic alteration, are recombined. In gene shuffling, a single sequence is mutated or otherwise altered and then recombined.

The generation and screening of high quality shuffled libraries provides for DNA shuffling (or "directed evolution"). The availability of appropriate high-throughput analytical chemistry to screen the libraries permits integrated high-throughput shuffling and screening of the libraries to achieve a desired activity.

In one significant embodiment, oligonucleotides for constructing codon-modified nucleic acids are designed in a computer ("in silico"). Predicted codon-modified recombinant nucleic acids can also be determined in silico, i.e., essentially as taught in Selifonov and Stemmer "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

Furthermore, rather than generating codon-modified nucleic acids as substrates for recombination, families of nucleic acids can be recombined simply by appropriate selection of the relevant oligonucleotides which are used in gene reconstruction methods to produce recombinant nucleic acids, i.e., by using codon-modified nucleic acid oligonucleotides as discussed herein in conjunction with family oligonucleotide-mediated shuffling methods, e.g., as taught in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. The technique can be used to recombine homologous or even non-homologous nucleic acid sequences; in the context of the present invention, oligonucleotides corresponding to families of codon-modified nucleic acids are shuffled.

The present invention provides significant advantages over previously used methods for optimization of genes. For example, DNA shuffling of codon modified nucleic acids can result in optimization of a desirable property even in the absence of a detailed

understanding of the mechanism by which the particular property is mediated. In addition, entirely new properties can be obtained upon shuffling of codon modified DNAs, i.e., shuffled DNAs can encode polypeptides or RNAs with properties entirely absent in the parental DNAs which are shuffled. Thus, by modifying the codon usage and/or encoded amino acids of the relevant gene or other nucleic acid, molecular diversity is accessed and sequences can be shuffled to obtain desired, including entirely new, properties.

In general, sequence recombination can be achieved in many different formats and permutations of formats, as described in further detail below.

The targets for modification vary in different applications, as does the property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have enzymatic or therapeutic or other commercially useful activities. A more extensive listing is found *supra*; however, even this list is not intended to be limiting, as essentially any nucleic acid can be codon modified and shuffled, using one or more of the processes herein.

Shuffling methods use at least two variant forms of a starting target (the variant forms can be nucleic acids, or representations thereof, e.g., as character strings in a computer program). The variant forms of candidate codon-altered substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one and preferably at least two positions. The initial diversity between forms can be the result of natural variation, *e.g.*, the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (*e.g.*, allelic variations), or constitute homologs from different organisms (interspecific variants), or constitute artificial homologs, *e.g.*, codon-altered nucleic acids encoding the same or a similar protein. Any or all of these sequences can represent or include codon altered nucleic acids.

Initial diversity can also be induced, *e.g.*, the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (*see,* Liao (1990) *Gene* 88:107-111), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further detail below, and are generally well known). The initial diversity between substrates is greatly augmented in subsequent steps of recombination for library generation.

A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism. The properties or characteristics that can be acquired or improved vary widely, and, of course depend on the choice of substrate.

At least two variant forms of a nucleic acid, e.g., which can confer a desired activity or which can be recombined to produce a desired activity, are recombined to produce a library of recombinant nucleic acids. The library is then screened to identify at least one recombinant nucleic acid that is optimized for the particular property or properties of interest.

Often, improvements are achieved after one round of recombination and selection. However, recursive sequence recombination can be employed to achieve still further improvements in a desired property, or to bring about new (or "distinct") properties. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity. That is, one creates a family of nucleic acid molecules showing some sequence identity to each other but differing due to the presence of mutations. In any given cycle, recombination can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying known methods of mutagenesis (*e.g.*, error-prone PCR or cassette mutagenesis) to either the substrates or products for recombination. In general, however, a single cycle of DNA shuffling of codon-altered nucleic acids provides for generation of surprisingly effective nucleic acids. Accordingly, while recursive approaches to shuffling can be used, single cycle recombination is also preferred. Typically, 2, 3, 4, 5, or even 10 or more cycles of recombination can be performed, each cycle optionally comprising one or more selection steps.

A recombination cycle is usually followed by at least one cycle of screening or selection for molecules having a desired property or characteristic. If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant segments, are sometimes introduced into cells before the screening step. Recombinant segments can also be linked to an appropriate vector or other regulatory sequences before screening.

Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (e.g., bacteriophage) before screening. If recombination is performed *in vivo*, recombination products can sometimes be screened in the cells in which recombination occurred. In other applications, recombinant segments are extracted from the cells, and optionally packaged as viruses, before screening.

The nature of screening or selection depends on what property or characteristic is to be acquired or the property or characteristic for which improvement is sought, and many examples are discussed below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have acquired new or improved properties or characteristics relative to the starting substrates. For example, a gene can have many component sequences, each having a different intended role (*e.g.*, coding sequences, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences can be varied and recombined simultaneously. Screening/selection can then be performed, for example, for recombinant segments that have increased ability to confer activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular screening protocol used for a desired property, initial round(s) of screening can sometimes be performed using bacterial cells due to high transfection efficiencies and ease of culture. However, bacterial expression is often not practical or desired, and yeast, fungal or other eukaryotic systems are also used for library expression and screening. Similarly, other types of screening which are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be performed in the precise cell type of intended use.

If further improvement in a property is desired, at least one and usually a collection of recombinant segments surviving a first round of screening/selection are subject to a further round of recombination. These recombinant segments can be recombined with each other or with exogenous segments representing the original substrates or further variants thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant segments as components of cells, the components can be

subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous screening step identifies desired recombinant segments in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates further recombinant segments which encompass additional diversity that is present in recombinant segments resulting from a previous round (or from multiple previous rounds, e.g., where the process is iteratively repeated).

The second round of recombination can be followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection can be increased between rounds. Also, the nature of the screen and the property being screened for can vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired. Additional rounds of recombination and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

The practice of this invention involves the construction of recombinant nucleic acids and the expression of genes in transfected host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the construction of recombinant nucleic acids such as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, including mutagenesis, include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel")). Methods of transducing cells, including plant and animal cells, with nucleic acids are generally available, as are methods of expressing proteins encoded by such nucleic acids. In addition to Berger, Ausubel and Sambrook, useful general references for culture of animal cells include Freshney (Culture of

Animal Cells, a Manual of Basic Technique, third edition Wiley- Liss, New York (1994)) and the references cited therein, Humason (Animal Tissue Techniques, fourth edition W.H. Freeman and Company (1979)) and Ricciardelli, et al., In Vitro Cell Dev. Biol. 25:1016-1024 (1989). References for plant cell cloning, culture and regeneration include Payne et al. (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York, NY (Payne); and Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) (Gamborg). A variety of Cell culture media are described in Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Additional information for plant cell culture is found in available commercial literature such as the Life Science Research Cell Culture Catalogue (1998) from Sigma- Aldrich, Inc (St Louis, MO) (Sigma-LSRCCC) and, e.g., the Plant Culture Catalogue and supplement (1997) also from Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

Examples of techniques sufficient to direct persons of skill through *in vitro* amplification methods, including the polymerase chain reaction (PCR), the ligase chain reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques (*e.g.*, NASBA) are found in Berger, Sambrook, and Ausubel, *id.*, as well as in Mullis *et al.*, (1987) U.S. Patent No. 4,683,202; PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1, 1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; Kwoh *et al.* (1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli *et al.* (1990) Proc. Natl. Acad. Sci. USA 87, 1874; Lomell *et al.* (1989) J. Clin. Chem 35, 1826; Landegren *et al.*, (1988) Science 241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene 4, 560; Barringer *et al.* (1990) Gene 89, 117, and Sooknanan and Malek (1995) Biotechnology 13: 563-564. Improved methods of cloning *in vitro* amplified nucleic acids are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng *et al.* (1994) Nature 369: 684-685 and the references therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. *See*, Ausbel, Sambrook and Berger, *all supra*.

Oligonucleotides *e.g.*, for use in *in vitro* amplification/ gene reconstruction methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method, e.g., as described by Beaucage and Caruthers (1981), *Tetrahedron Letts.*, 22(20):1859-1862, *e.g.*, using an automated synthesizer, e.g., as described in Needham-VanDevanter *et al.* (1984) *Nucleic Acids Res.*, 12:6159-6168 or as is now practiced routinely in the art. Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill. Purification of oligonucleotides (e.g., using gel-purification methods) to improve the quality of synthesized oligonucleotides can be particularly desirable in the processes herein to improve the quality of nucleic acid synthesis protocols.

As noted, essentially any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (http://www.genco.com), ExpressGen Inc. (www.expressgen.com), Operon Technoloigies Inc. (Alameda, CA) and many others. Similarly, peptides and antibodies can be custom ordered from any of a variety of sources, such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (http://www.htibio.com), BMA Biomedicals Ltd (U.K.), Bio Synthesis, Inc., and many others.

CODON AND AMINO ACID ALTERED LIBRARIES

In the methods of the invention, libraries of codon altered nucleic acids can be made and recombined. The codon altered nucleic acids can also include differences in encoded amino acid sequences, which can be either conservative or non-conservative in nature. The codon altered nucleic acids can be derived from a single parental amino acid sequence, or can be derived from a family of original sequences, *e.g.*, natural or synthetic homologous variants of a given sequence. Libraries can exist, *e.g.*, in pools or aliquots of cells, viral plaques, enzymatically synthesized pools or aliquots of nucleic acids, or chemically synthesisized pools of nucleic acids. Methods of making libraries of nucleic acids are available and taught, *e.g.*, in Berger, Sambrook and Ausubel, *supra*. In one embodiment, a library as used in the invention comprises at least 2 nucleic acid sequences. In additional

embodiments, the libraries of this invention comprise at least 2, 5, 10, 100, 1000, or more nucleic acid sequences.

As applied to the invention, libraries are typically constructed with a high percentage of codons altered relative to an initial (*e.g.*, wild type) nucleic acid. Codon usage divergence for each of the codon altered nucleic acids can be 50%, 75%, or even 90% or more as compared to the first nucleic acid. This eliminates hybridization to the parental nucleic acid (and thereby inhibits recombination with the parental nucleic acid, a desirable feature in certain embodiments discussed below).

In several embodiments of this invention, codons are modified in members of a gene family so as to increase the degree of identity between the members. In one such embodiment, the genes are homologous genes from different species. In such cases, the degree of nucleic acid identity may be lower than the degree of amino acid identity, at least in part, because of differences in codon usage between the species. In additional embodiments, the homologous genes represent different members of a gene family within a single species. Such genes may encode functionally distinct members of a gene family that nevertheless share significant structural or functional similarity. In preferred embodiments, homologous genes are reverse translated into nucleic acid sequences, and the nucleic acid sequences are modified so as to increase the level of identity between them. Nucleic acids with the modified sequences can then be synthesized *in vitro*. In particularly preferred embodiments, the modified nucleic acid sequences are at least as identical to each other as the original amino acid sequences.

Additional sequence diversity is provided by generating nucleic acids with non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid. This provides for reversion to wild-type upon recombination, while optionally allowing for the incorporation of non-conservative changes to the sequence in the event that they produce a detectable improvement during screening.

Modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide is also provided. This modification of core amino acids provides minor differences in encoded proteins, while changing the mutational spectrum of the resulting nucleic acid, thereby increasing sequence diversity.

In addition, due to the constraints of the translational machinery for a given cell, codon usage may need to be altered when expressed sequences are shuttled between different organisms (e.g., animal cells, plant cells, bacterial cells, etc.) for optimal expression. This produces a nucleic acid which encodes the same protein, but which, after typical forms of point mutation, will access a different mutational diversity than the original form of the protein.

In one embodiment, phage libraries are made and recombined in mutator strains such as cells with mutant or impaired gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. High multiplicity of infection (MOI) libraries are used to infect the cells to increase recombination frequency. Additional strategies for making phage libraries and or for recombining DNA from donor and recipient cells are set forth in U.S. Pat. No. 5,521,077. Additional recombination strategies for recombining plasmids in yeast are set forth in WO 97 07205.

The library to be made can be an *in vitro* set of molecules, or present in cells, phage or the like. Virtual libraries of nucleic acids generated in silico are also a feature of the invention (*see also*, Selifonov and Stemmer, *supra*). Generally, the library is screened to identify at least one recombinant nucleic acid that exhibits distinct or improved activity compared to the parental nucleic acid or nucleic acids which are recombined. Additional details on making appropriate libraries are found below, *e.g.*, in the section entitled "Formats for Sequence Recombination."

TARGETS FOR CODON MODIFICATION AND SHUFFLING

Essentially any nucleic acid can be codon altered and shuffled. No attempt is made herein to identify the hundreds of thousands of known nucleic acids. Common sequence repositories for known proteins include GenBank EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

One class of preferred targets for activation includes nucleic acids encoding therapeutic proteins such as erythropoietin (EPO), insulin, peptide hormones such as human growth hormone; growth factors and cytokines such as epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1β, MCP-1, epidermal growth

factor, fibroblast growth factor, hepatocyte growth factor, insulin-like growth factor, the interferons, the interleukins, keratinocyte growth factor, leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF etc. Many of these proteins are commercially available (*See,* e.g., the Sigma BioSciences 1997 catalogue and price list), and the corresponding genes are well-known.

Another class of preferred targets are transcriptional and expression activators. Example transcriptional and expression activators include genes and proteins that modulate cell growth, differentiation, regulation, or the like. Expression and transcriptional activators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, and animals, including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, e.g., by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and degrading RNA. Expression activators include cytokines, inflammatory molecules, growth factors, their receptors, and oncogene products, e.g., interleukins (e.g., IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, e.g., Mos, Ras, Raf, and Met; and transcriptional activators and suppressors, e.g., p53, Tat, Fos, Myc, Jun, Myb, Rel, and steroid hormone receptors such as those for estrogen, progesterone, testosterone, aldosterone, the LDL receptor ligand and corticosterone.

Similarly, proteins from infectious organisms for possible vaccine applications, described in more detail below, including infectious fungi, e.g., *Aspergillus*, *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (e.g., *aureus*), *Streptococci* (e.g., *pneumoniae*), *Clostridia* (e.g., *perfringens*), *Neisseria* (e.g., *gonorrhoea*), *Enterobacteriaceae* (e.g., *coli*), *Helicobacter* (e.g., *pylori*), *Vibrio* (e.g., *cholerae*), *Capylobacter* (e.g., *jejuni*), *Pseudomonas* (e.g., *aeruginosa*), *Haemophilus* (e.g., *influenzae*), *Bordetella* (e.g., *pertussis*), *Mycoplasma* (e.g., *pneumoniae*), *Ureaplasma* (e.g., *urealyticum*), *Legionella* (e.g., *pneumophila*), *Spirochetes* (e.g., *Treponema*, *Leptospira,* and *Borrelia*),

Mycobacteria (e.g., *tuberculosis*, *smegmatis*), *Actinomyces* (e.g., *israelii*), *Nocardia* (e.g., *asteroides*), *Chlamydia* (e.g., *trachomatis*), *Rickettsia*, *Coxiella*, *Ehrilichia*, *Rochalimaea*, *Brucella*, *Yersinia*, *Fracisella*, and *Pasteurella*; protozoa such as sporozoa (e.g., *Plasmodia*), *rhizopods* (e.g., *Entamoeba*) and flagellates (*Trypanosoma*, *Leishmania*, *Trichomonas*, *Giardia*, etc.); viruses such as (+) RNA viruses (examples include Poxviruses e.g., *vaccinia*; Picornaviruses, e.g. *polio*; Togaviruses, e.g., *rubella*; Flaviviruses, e.g., HCV; and Coronaviruses), (-) RNA viruses (examples include Rhabdoviruses, e.g., VSV; Paramyxovimses, e.g., RSV; Orthomyxovimses, e.g., influenza; Bunyaviruses; and Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, i.e., Retroviruses, e.g., especially HIV and HTLV, and certain DNA to RNA viruses such as Hepatitis B virus.

Other proteins relevant to non-medical uses, such as inhibitors of transcription or toxins of crop pests e.g., insects, fungi, weed plants, and the like, are also preferred targets for shuffling. Industrially important enzymes such as monooxygenases, proteases, nucleases, and lipases are also preferred targets. As an example, subtilisin can be evolved by shuffling codon altered forms of the gene for subtilisin (von der Osten et al., J. Biotechnol. 28:55-68 (1993) provide a subtilisin coding nucleic acid). Proteins which aid in folding such as the chaperonins are also preferred.

Preferred known genes suitable for codon alteration and shuffling also include the following: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, Apolipoprotein, Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Atrial peptides, C-X-C chemokines (e.g., T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4, SDF-1, PF4, MIG), Calcitonin, CC chemokines (e.g., Monocyte chemoattractant protein-1, Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309, R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, Collagen, Colony stimulating factor (CSF), Complement factor 5a, Complement inhibitor, Complement receptor 1, Factor IX, Factor VII, Factor VIII, Factor X, Fibrinogen, Fibronectin, Glucocerebrosidase, Gonadotropin, Hedgehog proteins (e.g., Sonic, Indian, Desert), Hemoglobin (for blood substitute; for radiosensitization), Hirudin, Human serum albumin, Lactoferrin, Luciferase, Neurturin, Neutrophil inhibitory factor (NIF), Osteogenic protein,

Parathyroid hormone, Protein A, Protein G, Relaxin, Renin, Salmon calcitonin, Salmon growth hormone, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, i.e., Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Toxic shock syndrome toxin (TSST-1), Exfoliating toxins A and B, Pyrogenic exotoxins A, B, and C, and M. arthritides mitogen, Superoxide dismutase, Thymosin alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha) and Urokinase. Many other known coding nucleic acids, such as those in Genebank™, can be codon-altered and shuffled.

GENES WITH CODON USAGE REDESIGNED AND CHEMICALLY SYNTHESIZED AS STARTING MATERIALS FOR GENE FAMILY SHUFFLING--EXPANDING THE DIVERSITY OF DNA SHUFFLING.

Because the genetic coding preference among organisms ranges from quite similar to very different, homologous genes from different organisms can have significantly lower homology at the nucleic acid level than at the amino acid level. For example, genetic information for some bacterial species is high in GC content (up to 70%), while others have AT rich (>60%) codon usage. Thus, genes from different organisms may have, for example, 40-60% amino acid identity but only 25-35% nucleic acid identity. It is often desirable to increase such levels of nucleic acid identity so as to enhance the ability of the homologous sequences to recombine, thereby increasing the efficiency of family shuffling using the methods of this invention. In other aspects, it is actually preferable to decrease the rate of recombination in a system, e.g., when using vectors it is sometimes desirable to decrease the rate of recombination between the vector and the host DNA, thereby increasing the safety of the vector. The following examples address specific issues with regard to shuffling codon altered nucleic acids.

Altering codon usage to increase homology

In one aspect, protein sequences of gene family members are reverse translated back into DNA sequences, for example by using one of the preferable codon usage charts in any conventional DNA manipulation program (*e.g.* the Wisconsin Package™, SeqWeb, OMIGA, SeqApp, SeqPup, MacVector, DNA stryder, GeneWorks, etc.). The

choice of codon usage is often determined by the host in which the genes will be expressed. After maximizing the percentage of DNA sequence identity, the genes are chemically synthesized, *e.g.*, using a high throughput oligonucleotide synthesizer in, *e.g.* a 96-well format, optionally in conjunction with polymerase and/ or ligase gene synthesis methods.

In general, the DNA sequence similarity after such treatment will be at least as high as the amino acid similarity, but can be at least about 10% to 15% *higher* than the amino acid identity (in contrast to the situation for naturally occurring genes, which are ordinarily less well conserved than encoded polypeptides), based on the random frequency of sequence identity for any given codon. In most cases, the minimal requirement for amino acid identity can be as low as about 35% while still retaining adequate nucleic acid homology for standard recombination methods (as discussed, *supra*, oligonucleotide-mediated recombination methods do not require high levels of similarity to achieve recombination). In some cases, however, the minimal amino acid identity can be even lower, *e.*g. if the conserved regions are clustered within the genes.

Example: Shuffling codon-modified EPO

The protein erythropoietin alpha, also known as EPO, Epogen, and Procrit is a hematopoietic hormone, providing a variety of benefits to patients suffering from anemia (a common symptom of, e.g., AIDS). EPO is produced as a pharmaceutical, with sales of nearly 1 billion dollars world-wide. Accordingly, proteins with EPO-like activity (and preferably superior activity) are of substantial commercial interest.

Figure 1 shows the sequence of a part of the monkey EPO gene, which is similar to the human EPO gene. Figure 2 shows an example of a codon altered EPO nucleic acid (or "wobble" EPO gene). In general, transversions rather than transition mutations are made where possible. The purpose of this strategy is to maximally disrupt hybridization of the resulting gene with naturally occurring EPOs. Figure 3 shows an alignment of naturally occurring EPOs.

This strategy is further fine-tuned by applying standard rules of base pairing (e.g., elimination of G-C pairing and GC stacking) to maximize sequence disruption; in addition, conservative or non-conservative amino acid modifications can also be made (in some cases, where multiple codon-altered nucleic acids are shuffled, it is desirable to make codon altered nucleic acids with non-overlapping non-conservative substitutions to permit

reversion to the wild-type amino acid during shuffling). The size of the sequence space for nucleic acids encoding EPO is large, at about 2.8 x 10^{88} different sequences (there are about 10^{80} particles in the universe; thus, it is physically impossible to make all of the possible sequences encoding EPO). As indicated schematically in Figure 4, if one only considers the maximally divergent wobble genes (those that use the alternative types of codons for leucine, arginine, and serine), there is still a sequence space of 10^{38} sequences encoding EPO. The overall strategy is to synthesize a library of wobble genes, screen for expression and activity and DNA shuffle desirable genes as desired (e.g., by recursive processes).

It is of interest to further evolve codon-altered nucleic acids. Shuffling with other homologous genes from nature, designed genes (incorporating libraries of designed sequence variation), and genes containing mutations of interest are strategies for evolving any gene of interest. However, the codon altered nucleic acid may not be easily shuffled with these genes because of the sequence differences; or they may be undesirable for other reasons (e.g., the naturally occurring sequences may be proprietary, or include proprietary elements). These difficulties can be avoided by synthesizing codon-altered homologous nucleic acids which encode desired amino acid variations (e.g., those found in homologous genes), but which have a codon-set close to the nucleic acid(s) to be recombined (thereby permitting, e.g., hybridization during recombination).

For example, after identifying homologues of interest (e.g., those shown in Figure 3 for EPO), codon-altered nucleic acids encoding the same proteins are synthesized with a similar codon selection. Standard family shuffling is then practiced with the codon altered nucleic acids. This is shown schematically for EPO in Figure 5.

EPO wobble variants are screened for expression and then receptor binding assays are conducted in an ELISA format, using human EPOr-Fc fusions. Following selection of binding variants, activity is measured as thymidine incorporation in UT7-EPO (A human bone marrow cell line) cell proliferation assays. Cells are treated for 2-3 days with various concentration of EPO variants after which time they are incubated in the presence of 3-H thymidine for 4 hours and incorporation of thymidine is measured. *See also*, Erickson-miller et al. (1997) Blood 90:2421 (for the receptor binding assay), and Wen et al. (1994) J. Biol. Chem. 269:22839-22846 (for the thymidine incorporation assay).

Assays for selecting EPO can also be based, e.g., on the ability of EPO proteins to stimulate the growth of blood cell, e.g., in vitro or in vivo.

Example: Codon Shuffling G-CSF

Family shuffling can be used to breed diversity from genes into the libraries to be screened. Additionally, design heuristics such as randomization of hydrophobic core residues can be used to take advantage of the redundancy between primary structure and tertiary structure of proteins (i.e. many different primary structures encode proteins with very similar three dimensional structures).

Design heuristics are employed to create a sequence space of mutants that are predicted to be highly biased (relative to random mutagenesis) to encode proteins which preserve the original activity. Methods such as high throughput (HTP) screening and phage panning are used to identify members of the designed libraries that have the desired activity. DNA shuffling is used to breed this population of active clones in order to fine tune the mutants, thus allowing one to evolve variants with equivalent or superior function relative to the naturally occurring proteins.

Figures 6 and 7 show several mammalian homologues of G-CSF. Figure 8 shows the hydrophobic core residues of human G-CSF (blacked out). Figure 9 shows a strategy for evolving variants of human G-CSF that are highly divergent in sequence. First, three genes are synthesized (Genes 1, 2 and 3, Figure 8) which contain all of the mammalian homologue diversity of G-CSF. These genes are shuffled, phage panned against the G-CSF receptor, and HTP screened for biological function (receptor activation). Active clones are iteratively shuffled and screened if necessary to give evolved variants that rival or surpass the human gene in activity (on human cells).

Next, one evolves a variant that has a highly mutated hydrophobic core. This is schematically illustrated in Figure 8, and the specific strategy for performing the biological screening is schematically illustrated in Figure 9. it is expected that the best mutants obtained after screening hydrophobic core randomized libraries may be less active than wild type human G-CSF because it is difficult to initially optimize activity in such a procedure. Family shuffling is used to obtain optimized variants. This is done by synthesizing genes which contain mammalian homologue diversity at all but the hydrophobic core positions; but

they are synthesized in the context of an evolved, non-wild type hydrophobic core. Family shuffling is used to optimize around the new hydrophobic core.

This strategy works because there are functionally similar hydrophobic cores for wild type proteins that consist of largely different amino acids than the wild type protein. This understanding is supported by recent experiments in model systems. For example, 53% of randomized sequences for three residues in the hydrophobic core of lambda repressor are folded and biologically active (Lim and Sauer (1991) J. Mol. Biol. 219:359-376). Protein design by patterning of polar and non-polar amino acids, where 24 residues in the hydrophobic core of a 4-helix bundle protein were randomized by Kamtekar et al. (1993) Science 262:1321. Folded, alpha helical proteins were recovered from about 1% of the clones. Desjarlais and Handel (1995) Current Opinion in Biotechnology 6:460-466 showed a mutant of Rop, another 4-helix bundle protein, where four hydrophobic core residues have been randomized and active mutants have been obtained. Axe et al. (1996) PNAS 95:5590-5594 showed that randomizing 13 hydrophobic core residues in the enzyme barnase resulted in 23% of the clones in the library retaining biological activity. Gassener et al. (1996) PNAS 93:12155-12158 describe a mutant of T4 lysozyme where 10 residues in the hydrophobic core are replaced with Met. This is taken as evidence that the hydrophobic core of this protein is very tolerant to substitution. Taken together, this experimental evidence on model systems shows that the hydrophobic cores of many proteins can be replaced with other hydrophobic residues that pack in a similar fashion to give an active protein. This degeneracy is exploited to evolve novel forms of natural and codon-altered genes.

A related approach is to search the protein databases for a protein that has a similar activity to a protein that on wishes to evolve. Denesyuk et al. (1996) J. Theor. Biol shows the results of such a search for G-CSF. LIF is a very similarly folded protein. One can use LIF as a `scaffold' on which to place residues of G-CSF that are required for activity. Given LIF with a G-CSF "toupee," one would family shuffle the LIF scaffold so as to obtain a variant in which the toupee is displayed in a fully biologically active form.

Another approach is to use computational methods to create families of variants that are predicted to be functional. Dahiyat and Mayo Science recently described computer methods that are used to design proteins. Proteins are simulated on the computer, often with the aid of genetic algorithms, and a subset that are deemed `fit' are actually

synthesized and 'analyzed'. These computational methods are becoming increasingly powerful. They would be useful to, for example, predict a family of mutations on the surface of a protein that would not destroy function. DNA shuffling can be used to optimized active clones obtained by design. Taking the example of G-CSF, one could use computational methods in combination with all structure function data (for example alanine scan data for G-CSF reported recently by Reidhaar-Olsen in Biochemistry) to design a family of putatively functional variants. One could, for example, design the family to have minimal DNA identity to the wild type gene given the design constraints. This library is synthesized, put through biological screens and/or selections (i.e. panning against the G-CSF receptor), and active variants are obtained. DNA shuffling is then used to evolve these active variants to have the desired level of function.

G-CSF proteins are displayed on phage and screened for binding to human G-CSF receptor in an ELISA format. Variants that bind receptor are selected in a high throughput screen for receptor activation. This cell based assay measures receptor activation via a reporter gene (such as luciferase) activated by a G-CSF responsive construct containing STAT binding elements. Cells (such as HepG2) are transformed with a G-CSF responsive reporter plasmid and treated with the codon shuffled G-CSF variant for 2.5 hours. Cells are then lysed and luciferase activity measured. *See also*, Tian et al (1998) Science 281:257-259.

Example: Codon Shuffling Alkaline Phosphatase

Alkaline phosphatase is a widely used reporter enzyme for ELISA assays, protein fusion assays, and in a secreted form as a reporter gene for mammalian cells. A more active form of the enzyme is desirable.

A codon altered form of alkaline phosphatase was generated by PCR assembly using the oligos set forth in Figure 10. A map of the oligos is set forth in Figure 11. The procedure used was essentially identical to that taught in Stemmer et al. (1994) Gene 164:49-57. In brief, the oligos were mixed 1:1 at a variety of dilutions and PCR assembled by performing e.g., 25-60 cycles of PCR at e.g., 94 °C (60 sec.), 94 °C (30 sec.), 50 °C (30 sec.), 72 °C (30 sec). Assembly of the BIAP gene was conducted in a circular format and gene fragements were purified. ~100,000 colonies were screened on LB/am plates (~1/10 are wt plasmid). About 1/10 showed a bluer color than background. Plasmid DNA showed a correct insertion.

In general, petri-dish screening using the typical colorimetric assay for phosphatase activity can be used for screening. This has the advantage or being simple, high throughput, and semi quantitative. Microtiter plate screening, also preferred, is colorimetric, and quantitative, although additional instrumentation can be required for implementation.

Example: Codon Shuffling to Reduce Competent Virus Production from Vectors and to Generate Attenuated Viruses as Immunogenic Compositions and Vaccines

Cells can be stably transduced with a number of viral vectors including those derived from retroviruses, pox viruses, adenoviruses (Ads), herpes viruses and parvoviruses. Common viral vectors include those derived from murine leukemia viruses (MuLV), gibbon ape leukemia viruses (GaLV), human immuno deficiency viruses (HIV), adenoviruses, adeno associated viruses (AAVs), Epstein Barr viruses, canarypox viruses, cowpox viruses, and vaccinia viruses. Viral vectors based upon retroviruses, adeno-associated viruses, herpes viruses and adenoviruses are all used as gene therapy vectors for the introduction of therapeutic nucleic acids into the cells of an organism by *ex vivo* and *in vivo* methods.

When using viral vectors, packaging cells are commonly used to prepare virions used to transduce target cells. In these vectors, trans-active genes are rendered inactive and "rescued" by trans-complementation to provide a packaged vector. This form of trans complementation is provided by co-infection of a packaging cell with a virus or vector which supplies functions missing from a particular gene therapy vector in trans, or by using a cell line (e.g., 293 cells) which have viral components integrated into the genome of the packaging cell. For instance, cells transduced with HIV or murine retroviral proviral sequences which lack the nucleic acid packaging site produce retroviral trans active components, but do not specifically incorporate the retroviral nucleic acids into the capsids produced, and therefore produce little or no live virus.

If these transduced "packaging" cells are subsequently transduced with a vector nucleic acid which lacks coding sequences for retroviral trans active functions, but includes a packaging signal, the vector nucleic acid is packaged into an infective virion. A number of packaging cell lines useful for MoMLV-based vectors are known in the art, such as PA317 (ATCC CRL 9078) which expresses MoMLV core and envelope proteins see, Miller et al. J. Virol. 65:2220-2224 (1991). Carrol et al. (1994) Journal of virology 68(9):6047-6051 describe the construction of packaging cell lines for HIV viruses.

Reciprocal complementation of defective HIV molecular clones is described, e.g., in Lori et al. (1992) Journal of Virology 66(9) 5553-5560.

Functions of viral replication not supplied by trans-complementation which are necessary for replication of the vector are present in the vector. In HIV, this typically includes, e.g., the TAR sequence, the sequences necessary for HIV packaging, the RRE sequence if the instability elements of the p17 gene of gag is included, and sequences encoding the polypurine tract. HIV sequences that contain these functions include a portion of the 5' long terminal repeat (LTR) and sequences downstream of the 5' LTR responsible for efficient packaging, i.e., through the major splice donor site ("MSD"), and the polypurine tract upstream of the 3' LTR through the U3R section of the 3' LTR. The packaging site (psi site or ψ site) is partially located adjacent to the 5' LTR, primarily between the MSD site and the gag initiator codon (AUG) in the leader sequence. See, Garzino-Demo et al. (1995) Hum. Gene Ther. 6(2): 177-184. For a general description of the structural elements of the HIV genome, see, Holmes et al. PCT/EP92/02787.

Another common vector is based upon adenovirus. Typically, vectors which include the adenovirus ITRs (Gingeras *et al.* (1982) *J. Biol. Chem.* 257:13475-13491) are packaged in, e.g., 293 cells, which provide many of the components necessary for vector packaging.

Adeno-associated viruses (AAVs) utilize helper viruses such as adenovirus or herpes virus to achieve productive infection. In the absence of helper virus functions, AAV integrates (site-specifically) into a host cell's genome, but the integrated AAV genome has no pathogenic effect. The integration step allows the AAV genome to remain genetically intact until the host is exposed to the appropriate environmental conditions (*e.g.*, a lytic helper virus), whereupon it re-enters the lytic life-cycle. Samulski (1993) *Current Opinion in Genetic and Development* 3:74-80 and the references cited therein provides an overview of the AAV life cycle. For a general review of AAVs and of the adenovirus or herpes helper functions *see*, Berns and Bohensky (1987) *Advanced in Virus Research*, Academic Press., 32:243-306. The genome of AAV is described in Laughlin *et al.* (1983) *Gene*, 23:65-73. Expression of AAV is described in Beaton *et al.* (1989) *J. Virol.*, 63:4450-4454. In general, the packaging sites for all parvoviruses, including B 19 and AAV are located in the viral ITRs. Recombinant AAV vectors (rAAV vectors) deliver foreign nucleic acids to a wide

range of mammalian cells (Hermonat & Muzycka (1984) *Proc Natl Acad Sci USA* 81:6466-6470; Tratschin *et al.* (1985) *Mol Cell Biol* 5:3251-3260), integrate into the host chromosome (McLaughlin *et al.* (1988) *J Virol* 62: 1963-1973), and show stable expression of the transgene in cell and animal models (Flotte *et al.* (1993) *Proc Natl Acad Sci USA* 90:10613-10617). rAAV vectors are able to infect non-dividing cells (Podsakoff *et al.* (1994) *J Virol* 68:5656-66; Flotte *et al.* (1994) *Am. J. Respir. Cell Mol. Biol.* 11:517-521). Further advantages of rAAV vectors include the lack of an intrinsic strong promoter, thus avoiding possible activation of downstream cellular sequences, and the vector s naked icosohedral capsid structure, which renders the vectors stable and easy to concentrate by common laboratory techniques.

One problem with previously existing vector packaging strategies is that vectors to be packaged can recombine with nucleic acids providing packaging functions in trans, producing a replication-competent virus. This can be a problem both when vectors are produced for therapeutic applications (e.g., in gene therapy) and during production of encoded components in vitro. The present invention provides a way of reducing or eliminating recombination between nucleic acids encoding trans-active components and vector nucleic acids encoding packaging sites.

In particular, nucleic acid subsequences of a vector which are adjacent to modified or deleted elements provided in trans, are codon modified to eliminate hybridization to wild-type sequences. Because these sequences do not hybridize, they cannot recombine with nucleic acids producing trans-active components. One additional advantage of this approach is that the vectors also cannot recombine with live viruses, e.g., in a human body which is infected with a virus that packages vector elements. As noted above, two types of gene therapy vectors are those based upon retroviruses (which can be packaged by, e.g., HIV-1) and adenoviruses (which can be packaged by adenovirus).

Alternatively, the nucleic acids encoding trans-active components can be codon modified so that they do not hybridize to wild-type sequences. This also prevents recombination with vectors having wild-type sequences, preventing recombination and formation of replication competent viruses.

After codon modification, vectors or trans active nucleic acids can be shuffled as described *supra*, and screened for the ability to package nucleic acids, or to be packaged, as appropriate.

It will be appreciated that codon modification of viral sequences has an additional use as well. Codon alteration of viral sequences can result in attenuation of the virus, e.g., due to modification of regulatory sequences, alterations in mRNA secondary structure, inefficient translation due to rare codon use, and the like. Such "codon attenuated" viruses have a significant advantage over existing attenuated viruses (which are typically generated by serial passage in cells other than the normal host type for the virus). In particular, codon attenuated viruses can encode a wild-type set of proteins, making them ideal as immunogenic compositions to generate antibodies, or to use as vaccines. Viral proteins can also be used in various diagnostic assays. For example, the standard diagnostic test for HIV infection in current use tests for the presence of anti-HIV antibodies in blood by probing with viral proteins.

Example: Codon Usage Libraries to evolve Functional Variants with reduced Recombination With Natural Gene Sequences--Adenovirus

Adenovirus is a common vector used, e.g., for gene therapy. The virus is typically modified to make it replication deficient. This can be achieved e.g., by deleting the E1 and E4 genes. The functions of E1 and E4 can be supplied by trans complementation when E1 and E4 deleted vectors are grown in the ubiquitous human embryonic kidney cell line 293, which has uncharacterized adenovirus fragments incorporated into their genome that supply the missing functions in trans. The replication defective adenoviral vectors recombine at a low, but clinically significant frequency, resulting in replication competent adenovirus contamination of vector preparations. Because adenovirus has detrimental effects on health, this is a significant problem for application of adenovirus-based gene therapy vectors.

In the present invention, a codon usage library encompassing several hundred bases to several kilobases of sequence flanking the adenovirus E1 and E4 genes are made. The library is designed to enforce a high degree of divergence from the natural adenoviral consensus sequence, while at the same time incorporating a large degree of degeneracy in the codons to allow for a large space of sequence diversity to be searched. The design principle

is to obtain mutants that encode the same or similar protein sequence, but with many mismatches to the wild-type E1 and E4 sequences found in the 293 genome. These mismatches strongly reduce the frequency of unwanted recombination with the trans complementary genes. Consequently, engineered adenoviral vectors, or adenovirus helper vectors which package adenoviral sequences which include packaging sequences (adenoviral or adeno-associated viral ITRs) in trans have reduced levels of recombination. This provides for a lower rate of competent adenovirus production, making culture and production of such vectors safer.

Evolution Impaired Viruses Created by Massive Codon Usage Alteration As a General Approach to Vaccines--HIV

HIV-1 and HIV-2 are genetically related, antigenically cross reactive, and share a common cellular receptor (CD4). *See*, Rosenburg and Fauci (1993) in *Fundamental Immunology, Third Edition* Paul (ed) Raven Press, Ltd., New York (Rosenburg and Fauci 1) and the references therein for an overview of HIV infection. HIV-1 infection is epidemic world wide, causing a variety of immune system-failure related phenomena commonly termed acquired immune deficiency syndrome (AIDS). HIV type 2 (HIV-2) has been isolated from both healthy individuals and patients with AIDS-like illnesses (Andreasson, *et al.* (1993) *Aids* 7, 989-93; Clavel, *et al.* (1986) *Nature*, 324, 691-695; Gao, *et al.* (1992) *Nature* 358, 495-9; Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Kanki, *et al.* (1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30; Naucler, *et al.* (1993) *International Journal of STD and Aids* 4, 217-21; Naucler, *et al.* (1991) *Aids* 5, 301-4). Although HIV-2 AIDS cases have been identified principally from West Africa, sporadic HIV-2 related AIDS cases have also been reported in the United States (O'Brien, *et al.* (1991) *Aids* 5, 85-8) and elsewhere. HIV-2 will likely become endemic in other regions over time, following routes of transmission similar to HIV-1 (Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30). Epidemiological studies suggest that HIV-2 produces human disease with lesser penetrance than HIV-1, and exhibits a considerably longer period of clinical latency (at least

25 years, and possibly longer, as opposed to less than a decade for HIV-1; *see*, Kanki, *et al.* (1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30, and Travers *et al.* (1995) *Science* 268: 1612-1615).

The ability of HIV virus populations to rapidly point mutate to avoid the immune response poses a special challenge for vaccine design. While the immune system has responded to viruses in a gradual and co-evolutionary manner, the present invention provides a general approach that provides for massively faster evolution to produce new vaccines to stimulate more effective immune responses.

For example, during the incubation period for HIV infection, which lasts for several years, low titers of HIV can result from high HIV replication rates in conjunction with efficient viral clearance by the immune system. In response to these selective forces, virus mutations are selected which reduce recognition and neutralization by the immune system's B and T-cell responses. *See,* Lukashov et al. (1995) J. Virol. 69:6911-6916. During the long incubation time, these mutations accumulate and eventually overwhelm the immune system's defenses. *See,* Ho et al. (1995) Nature.

Live attenuated vaccines, typically produced by prolonged growth of human viruses in animal cells, have proven useful as vaccines for several diseases, including mumps, rubella and measles. Attenuation involves the slow accumulation of many mutations throughout the viral genome during the course of adapting to growth in the animal cells. When used to vaccinate humans, the attenuated virus grows only weakly and elicits a complex immune response which the virus is unable to avoid. The mutations in the attenuated virus could, in principle, revert in the same stepwise fashion that it underwent to grow in culture.

The risk of reversion is highest in viruses with a high mutation rate such as HIV-1, which makes this strategy dangerous under current techniques for vaccine development. It is worth noting, however, that protective effects against HIV-1 are observed following infection with the related HIV-2 virus, which is much less pathogenic than HIV-1. Thus, protective effects against HIV can be achieved with live vaccines.

To reduce the risk of reversion, a large number of mutations need to accumulate in the virus. However, if too many mutations are present, the immune system in

effect recognizes the attenuated virus, but not the virus against which a protective effect is sought.

As provided herein, immunogenic compositions such as vaccines are created which contain a large number of silent substitutions. In contrast to existing attenuated viruses, such viruses have native protein sequences and elicit essentially the same immune responses as the corresponding wild-type virus (typically one or a few additional disabling mutations can also be incorporated). Codon alteration results in two effects that both increase the potential of the vaccine.

First, like standard attenuated viruses, the growth of codon-altered viruses is attenuated, due to the effect of the codon alterations on translation, regulatory sequences, mRNA folding, packaging, and the like. For example, regulation of HIV-1 envelope expression has been observed as a result of codon usage. *See,* Haas et al. (1989) Current Biology 6(3):315-324.

Second, codon alteration results in impairment of virus evolution. As discussed above, modification of the codons alters the mutational escape spectrum of the virus, upsetting the evolutionary selection for specific codons.

The six codon amino acids are the best targets for codon alteration. Serine, arginine and leucine each have one group of four codons, plus two codons in an unrelated group. *See*, Figure 12. Switching all of the serine codons from AGY to TCX and vice versa, yields proteins with unaltered amino acid sequences. *See also*, Figure 13. However, these codon groups differ significantly in the spectrum of the amino acids that they yield upon point mutation. Of all possible point mutations of one codon for serine (TCA) 78% result in a different amino acid compared to point mutations obtained for the AGT codon for serine. *See*, Figure 13. A virus with hundreds of codon alterations is in, statistically, a very different mutational space, able to access a totally different mutation spectrum, or "cloud," compared to the wild-type virus. The overall strategy for producing an evolution-defective virus is additionally set forth in Figures 14, 15 and 17. Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Point mutation is critical for viruses such as HIV-1 to stay ahead of the host immune system. The amino acid mutations that are required for virus escape are likely not random. Wild type codon usage has evolved to allow optimal immune system evasion. The

wild type codon usage is likely to favor mutations that represent alterations that avoid the host immune system, without detrimentally affecting the protein(s) encoded. While complex, this natural pattern of amino acid sequence change of the natural virus in response to the host system is non-random and weakly predictable. *See also*, Seiller-Moiseiwitsch et al. (1994) Annu. Rev. Genet. 28:559-596.

Changing all of the codons for ser, arg and leu in the 875 aa envelope polyprotein of HIV-1 (e.g., strain MN) would affect 187 codons (22%) resulting in 561 mutations. *See*, e.g., Figure 18, panels A-D. If all of the HIV proteins were altered, the number of mutations would be more than three-fold higher. The construction of such codon modified viruses is simplified by recent advances in the synthesis of long DNA sequences, which enable the assembly of a plasmid of average size from 40 mer oligos in a single step with about 75% efficiency. *See*, Stemmer (1994) Nature 370_389-391. *See also*, Figure 19 for a list of oligos in one application for synthesis of HIV Env. While the synthesis of the envelope gene is sufficient, synthesis of the whole HIV genome from oligos can be performed by this method.

In practice, a preferred balance of attenuation and evolution impairment is obtained by DNA shuffling (e.g., Stemmer et al. (1995) Gene 164:49-53), e.g., of the wild-type and codon altered sequences, followed by selection of the resulting library of viruses that retain moderate growth despite many codon alterations.

While attenuation that can be obtained by this approach may be sufficient for obtaining a vaccine for most viruses, for HIV-1, the evolution impairment is more important, due to the high mutation rate of the virus. Live vaccines are used only if they elicit an immune response which is complex and strong enough to prevent infection of the wild-type virus. Live virus vaccines are typically more protective than single protein vaccines because it is harder to out-mutate T and B-cell responses to a larger number of epitopes. The weak growth of the live virus vaccine results in a larger antigenic dose and point mutation is increases the complexity of the immune response. To evaluate vaccine competence, vaccine potential is evaluated in Macaques (*M. nemestrina*) or chimpanzees using SIV variants that are known to cause AIDS. Sequence for an example SIV, SIVsmm, is found at Gene Bank Accession No. x14307. This virus is closely related to HIV-2. *See*, Hirsch (1989) Nature 339: 389-392. In general, many complete sequences for HIVs, SIVs and many other viruses

are found in well known sequence repositories, including GenBank, EMBL, DDBJ and the NCBI. Well characterized HIV clones include: HIV-1NL43, HIV-1SF2, HIV-1BRU and HIV-1MN. For an introduction to the genetic variability of HIV, *see,* Seillier-Moiseiwitsch et al. (1994) Annu. Rev. Genet. 28:559-96 and the references cited therein.

Several HIV-2 isolates, including three molecular clones of HIV-2 ($HIV\text{-}2_{ROD}$, $HIV\text{-}2_{SBL\text{-}ISY}$, and $HIV\text{-}2_{UCI}$), have also been reported to infect macaques (*M. mulatta* and *M. nemestrina*) or baboons (Franchini, *et al.* (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2433-2437; Barnett, *et al.* (1993) *Journal of Virology* 67, 1006-14; Boeri, *et al.* (1992) *Journal of Virology* 66, 4546-50; Castro, *et al.* (1991) *Virology* 184, 219-26; Franchini, *et al.* (1990) *Journal of Virology* 64, 4462-7; Putkonen, *et al.* (1990) *Aids* 4, 783-9; Putkonen, *et al.* (1991) *Nature* 352, 436-8). As human pathogens capable of infection of small primates, HIV-2 molecular clones provide attractive models for studies of AIDS pathogenesis, and for drug and vaccine development against HIV-1 and HIV-2.

Recently, HIV-2 was suggested as a possible vaccine candidate against the more virulent HIV-1 due to its long asymptomatic latency period, and its ability to protect against infection by HIV-1 (*see*, Travers *et al.* (1995) *Science* 268: 1612-1615 and related commentary by Cohen *et al* (1995) *Science* 268: 1566). In the nine-year study by Travers *et al.* (*id*) of West African prostitutes infected with HIV-2, it was determined that infection with HIV-2 caused a 70% reduction in infection by HIV-1. Thus, codon altered HIV-2 viruses can also be used as a live vaccine, against both HIV-2 and HIV-1. Furthermore, because the natural pathogenicity of HIV-2 is less than HIV-1, it is, in addition to HIV-1, a preferred virus for modification.

FORMATS FOR SEQUENCE RECOMBINATION

The methods of the invention entail performing recombination ("shuffling") and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer (1995) *Bio/Technology* 13:549-553). Reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest. Such techniques do not require the extensive analysis and computation required by conventional methods for polypeptide engineering. Shuffling allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to natural pair-wise recombination events (e.g., as occur during sexual

replication). Thus, the sequence recombination techniques described herein provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result. In some instances, however, structural and/or functional information is available which, although not required for sequence recombination, provides opportunities for modification of the technique.

Exemplary formats and examples for sequence recombination, referred to, e.g., as "DNA shuffling," "fast forced evolution," or "molecular breeding," have been described by the present inventors and co-workers in the following patents and patent applications: US Patent No. 5,605,793; PCT Application WO 95/22625 (Serial No. PCT/US95/02126), filed February 17, 1995; US Serial No. 08/425,684, filed April 18, 1995; US Serial No. 08/621,430, filed March 25, 1996; PCT Application WO 97/20078 (Serial No. PCT/US96/05480), filed April 18, 1996; PCT Application WO 97/35966, filed March 20, 1997; US Serial No. 08/675,502, filed July 3, 1996; US Serial No. 08/721, 824, filed September 27, 1996; PCT Application WO 98/13487, filed September 26, 1997; "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination" Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre et al. (PCT/US99/15972, filed 07/15/1999); Stemmer, *Science* 270:1510 (1995); Stemmer *et al.*, *Gene* 164:49-53 (1995); Stemmer, *Bio/Technology* 13:549-553 (1995); Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* 91:10747-10751 (1994); Stemmer, *Nature* 370:389-391 (1994); Crameri *et al.*, *Nature Medicine* 2(1):1-3 (1996); and Crameri *et al.*, *Nature Biotechnology* 14:315-319 (1996), each of which is incorporated by reference in its entirety for all purposes.

The recombination procedure starts with at least two substrates that generally show substantial sequence identity to each other (*e.g.*, at least about 30%, 50%, 70%, 80% or 90% or more sequence identity), but differ from each other at certain positions. For example, at least one codon altered nucleic acid is recombined with one or more additional nucleic acid (the additional nucleic acid can also be a codon altered nucleic acid) herein. The difference between nucleic acids to be recombined can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in about 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide

positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. More typically, they will be codon altered nucleic acids derived from one or more homologous nucleic acid sequence. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness (*e.g.*, codon altered nucleic acids derived from different, but homologous, genes within a superfamily). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example; one segment can be a subsequence of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form.

The starting DNA segments are recombined by any of the sequence recombination formats provided herein to generate a diverse library of recombinant DNA segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5, 10^9, 10^{12}, 10^{15}, 10^{20} or even more members. In some embodiments, the starting segments and the recombinant libraries generated will include essentially full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. In other embodiments, the recombinant DNA segments in the library can be inserted into a common vector providing sequences necessary for expression before performing screening/selection.

Use of Restriction Enzyme Sites to Recombine Mutations

In some situations it is advantageous to use restriction enzyme sites in nucleic acids to direct the recombination of mutations in a nucleic acid sequence of interest. These techniques are particularly preferred in the evolution of fragments that cannot readily be shuffled by other existing methods due to the presence of repeated DNA or other problematic

primary sequence motifs. These situations also include recombination formats in which it is preferred to retain certain sequences unmutated. The use of restriction enzyme sites is also preferred for shuffling large fragments (typically greater than 10 kb), such as gene clusters that cannot be readily shuffled and "PCR-amplified" because of their size. Although fragments up to 50 kb have been reported to be amplified by PCR (Barnes, *Proc. Natl. Acad. Sci. U.S.A.* 91:2216-2220 (1994)), it can be problematic for fragments over 10 kb, and thus alternative methods for shuffling in the range of 10 - 50 kb and beyond are preferred. Preferably, the restriction endonucleases used are of the Class II type (Sambrook, Ausubel and Berger, *supra*) and of these, preferably those which generate nonpalindromic sticky end overhangs such as Alwn I, Sfi I or BstX1. These enzymes generate nonpalindromic ends that allow for efficient ordered reassembly with DNA ligase. Typically, restriction enzyme (or endonuclease) sites are identified by conventional restriction enzyme mapping techniques (Sambrook, Ausubel, and Berger, *supra.*), by analysis of sequence information for that gene, or by introduction of desired restriction sites into a nucleic acid sequence by synthesis (*i.e.* by incorporation of silent mutations). For example, one or more codon-altered nucleic acid can be recombined at restriction sites, e.g., with one or more nucleic acid of interest (including, e.g. a gene or gene cluster to be modified by recombination with the codon-altered nucleic acid).

The DNA substrate molecules to be digested can either be from *in vivo* replicated DNA, such as a plasmid preparation, or from synthetic or e.g., PCR amplified nucleic acid fragments harboring the restriction enzyme recognition sites of interest, preferably near the ends of the fragment. Typically, at least two variants of a gene of interest, each having one or more mutations, and at least one of which incorporating codon-modifications, are digested with at least one restriction enzyme determined to cut within the nucleic acid sequence of interest. The restriction fragments are then joined with DNA ligase to generate full length genes having shuffled regions. The number of regions shuffled will depend on the number of cuts within the nucleic acid sequence of interest. The shuffled molecules can be introduced into cells as described above and screened or selected for a desired property as described herein. Nucleic acid can then be isolated from pools (libraries), or clones having desired properties and subjected to the same procedure until a desired degree of improvement is obtained.

In some embodiments, at least one DNA substrate molecule or fragment thereof is isolated and subjected to mutagenesis. In some embodiments, the pool or library of religated restriction fragments are subjected to mutagenesis or additional recombination protocols before the digestion-ligation process is repeated. "Mutagenesis" as used herein comprises such techniques known in the art as PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, etc., and recursive sequence recombination by any of the techniques described herein.

Reassembly PCR

A further technique for recombining mutations in a nucleic acid sequence utilizes "reassembly PCR." This method can be used to assemble multiple segments that have been separately evolved into a full length nucleic acid template such as a gene. This technique is performed when a pool of advantageous mutants is known from previous work, or has been identified by screening mutants that may have been created by any mutagenesis technique known in the art, such as PCR mutagenesis, cassette mutagenesis, doped oligo mutagenesis, chemical mutagenesis, or propagation of the DNA template *in vivo* in mutator strains. Boundaries defining segments of a nucleic acid sequence of interest preferably lie in intergenic regions, introns, or areas of a gene not likely to have mutations of interest. Preferably, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of the nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length.

Each of the segments is amplified with a set of such primers. The PCR products are then "reassembled" according to assembly protocols such as those discussed herein to assemble randomly fragmented genes. In brief, in an assembly protocol the PCR products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTPs) and appropriate buffer salts in the absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes. In some embodiments, the resulting reassembled genes are subjected to mutagenesis before the process is repeated.

In the present invention, oligos such as PCR primers can include codon modifications as compared to a starting sequence. In addition, oligonucleotides can form the basis for PCR concatemerization reactions in which overlapping hybridized oligonucleotides are extended in one or more PCR amplification cycles. In this embodiment, a template nucleic acid is not required (although a template or fragments thereof can be added to the amplification mixture, which can aid in the eventual reassembly of a full-length gene). Further details regarding oligonucleotide gene reassembly methods are found, e.g., in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049.

In a further embodiment, the PCR primers for amplification of segments of a nucleic acid sequence of interest are used to introduce variation into the gene of interest as follows. Mutations at sites of interest in a nucleic acid sequence are identified by screening or selection, by sequencing homologues of the nucleic acid sequence, and so on. Oligonucleotide PCR primers are synthesized which encode wild type or mutant information at sites of interest. These primers are then used in PCR mutagenesis to generate libraries of full length genes encoding permutations of wild type and mutant information at the designated positions. This technique is typically advantageous in cases where the screening or selection process is expensive, cumbersome, or impractical relative to the cost of sequencing the genes of mutants of interest and synthesizing mutagenic oligonucleotides.

Site Directed Mutagenesis (SDM) with Oligonucleotides Encoding Homologue Mutations Followed by Shuffling

In some embodiments of the invention, sequence information from one or more substrate sequences is added to a given "parental" sequence of interest, with subsequent recombination between rounds of screening or selection. Typically, this is done with site-directed mutagenesis performed by techniques well known in the art (e.g., Berger, Ausubel and Sambrook, *supra*.) with one substrate as a template and oligonucleotides encoding single or multiple mutations from other substrate sequences, *e.g.* homologous genes. After screening or selection for an improved phenotype of interest, the selected recombinant(s) can be further evolved using recursive techniques. After screening or selection, site-directed mutagenesis can be done again with another collection of oligonucleotides encoding

homologue mutations, and the above process repeated until the desired properties are obtained.

When the difference between two homologues is one or more single point mutations in a codon, degenerate oligonucleotides can be used that encode the sequences in both homologues. One oligonucleotide can include many such degenerate codons and still allow one to exhaustively search all permutations over that block of sequence.

When the homologue sequence space is very large, it can be advantageous to restrict the search to certain variants. Thus, for example, computer modeling tools (Lathrop *et al.* (1996) *J. Mol. Biol.*, 255: 641-665) can be used to model each homologue mutation onto the target protein and discard any mutations that are predicted to grossly disrupt structure and function. In silico genetic algorithm operations for generating and predicting mutational events are found in Selifonov and Stemmer "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

Oligonucleotide and in silico shuffling formats

As mentioned above, at least two additional related formats are useful in the practice of the present invention. The first, referred to as "in silico" shuffling utilizes computer algorithms to perform "virtual" shuffling using genetic operators in a computer. As applied to the present invention, codon altered gene sequence strings are recombined in a computer system and desirable products are made, e.g., by reassembly PCR or ligation of synthetic oligonucleotides. In silico shuffling is described in detail in Selifonov and Stemmer in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854. In brief, genetic operators (algorithms which represent given genetic events such as point mutations, recombination of two strands of homologous nucleic acids, etc.) are used to model recombinational or mutational events which can occur in one or more nucleic acid, e.g., by aligning nucleic acid sequence strings (using standard alignment software, or by manual inspection and alignment) and predicting recombinational outcomes. The predicted recombinational outcomes are used to produce corresponding molecules, e.g., by oligonucleotide synthesis and reassembly PCR.

The second useful format is referred to as "oligonucleotide mediated shuffling" in which oligonucleotides corresponding to a family of related homologous nucleic acids (e.g., as applied to the present invention, codon modified synthetic homologous variants of a nucleic acid) which are recombined to produce selectable nucleic acids. This format is described in detail in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. In brief, selected oligonucleotides are synthesized, ligated and elongated, typically either in a polymerase or ligase-mediated elongation reaction. The technique can be used to recombine homologous or even non-homologous codon-altered nucleic acid sequences.

One advantage of oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of fragmented nucleic acids (e.g., cleaved codon-modified oligonucleotides, or synthesized codon-modified oligonucleotides) are recombined, e.g., with a with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination.

When recombining homologous nucleic acids, sets of overlapping family gene shuffling oligonucleotides (which are derived by comparison of homologous nucleic acids that include one or more codon-modified nucleic acid, followed by synthesis of corresponding oligonucleotides) are hybridized and elongated (e.g., by reassembly PCR or ligation), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. The set of overlapping family shuffling gene oligonucleotides includes a plurality of oligonucleotide member types which have consensus region subsequences derived from a plurality of homologous target nucleic acids.

Typically, as applied to the present invention, family gene shuffling oligonucleotide which include one or more codon-altered nucleic acid(s) are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are synthesized (serially or in parallel) which correspond to at least one region of sequence diversity.

Sets of fragments, or subsets of fragments used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (e.g., with a DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments can be used in conjunction with family gene shuffling oligonucleotides, e.g., in one or more recombination reaction to produce recombinant codon-altered nucleic acid(s).

Additional oligonucleotide shuffling formats are found in co-filed application by Crameri et al., "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" (Attorney Docket Number 02-296-2US) and in co-filed application by Welch et al., "USE OF CODON VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" (Attorney docket number 02-1007). In particular, these applications provide for tri-nucleotide-based synthesis of degenerate oligonucleotides, thereby providing for codon substitution during oligonucleotide shuffling. In brief, this procedure utilizes tri-nucleotide phosphoramidite chemistry to synthesize oligos, rather than standard mono-nucleotide synthesis. Because codons are altered as a unit, the synthetic scheme of degenerate oligonucleotides is simplified.

Additional In Vitro DNA Shuffling Formats

In one embodiment for shuffling DNA sequences *in vitro*, the initial substrates for recombination are a pool of related sequences, *e.g.*, different variant forms, as homologs from different individuals, strains, or species of an organism, or related sequences from the same organism, as allelic variations. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or reassembled. Preferably the sequences are from 50 base pairs (bp) to 50 kilobases (kb).

The pool of related substrates are converted into overlapping fragments, *e.g.*, from about 5 bp to 5 kb or more. Often, for example, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNase I or RNase digestion, random shearing or partial restriction enzyme digestion, or by oligonucleotide synthesis as in the family oligonucleotide-mediated shuffling methods of crameri et al., discusses *supra*. For discussions of protocols for the isolation, manipulation, enzymatic digestion, and the like, of nucleic acids, *see*, for example, Sambrook *et al.* and Ausubel, both *supra*. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 2, 10, 100, 500 or 1,000 or more.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form using any of a variety of techniques, including, for example, heating, chemical denaturation, use of DNA binding proteins, and the like (in oligonucleotide mediated methods, this step can be omitted). Conversion can be effected by heating to about 80 °C to 100 °C, more preferably from 90 °C to 96 °C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with a single-stranded DNA binding protein (see Wold (1997) *Annu. Rev. Biochem.* 66:61-92) or *rec*A protein (*see, e.g.*, Kiianitsa (1997) *Proc. Natl. Acad. Sci. U S A* 94:7837-7840). Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 20 °C to 75 °C, and preferably from 40 °C to 65 °C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can

be used with an annealing temperature of between 20-30 °C. The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase or ligase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as shuffling of the nucleic acid *in vitro*. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from shuffling is used to transform host cells, optionally after cloning into a vector.

In one embodiment utilizing *in vitro* shuffling, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. Another embodiment uses random primers to prime an entire template DNA to generate less than full length amplification products. The amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, in which at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the partially extended (less than full length) products reanneal to and prime extension on different sequence-related template species. In another embodiment, the conversion of substrates to fragments can be effected by partial PCR amplification of substrates.

In another embodiment, a mixture of fragments is spiked with one or more oligonucleotides. The oligonucleotides can be designed to include precharacterized mutations of a wildtype sequence (e.g., codon modification), or sites of natural variations between individuals or species. The oligonucleotides also typically include sufficient sequence or structural homology flanking such mutations or variations to allow annealing

with the wildtype fragments. Annealing temperatures can be adjusted depending on the length of homology.

In a further embodiment, recombination occurs in at least one cycle by template switching, such as when a DNA fragment derived from one template primes on the homologous position of a related but different template. Template switching can be induced by addition of recA (*see*, Kiianitsa (1997) *supra*), rad51 (*see*, Namsaraev (1997) *Mol. Cell. Biol.* 17:5359-5368), rad55 (*see*, Clever (1997) *EMBO J.* 16:2535-2544), rad57 (*see*, Sung (1997) *Genes Dev.* 11:1111-1121) or other polymerases (*e.g.*, viral polymerases, reverse transcriptase) to the amplification mixture. Template switching can also be increased by increasing the DNA template concentration.

Another embodiment utilizes at least one cycle of amplification, which can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Fragments can be prepared using a single stranded DNA phage, such as M13 (*see*, Wang (1997) *Biochemistry* 36:9486-9492). Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, ssDNA fragments of variable length can be generated from a single primer by Pfu, Taq, Vent, Deep Vent, UlTma DNA polymerase or other DNA polymerases on a first DNA template (*see*, Cline (1996) *Nucleic Acids Res.* 24:3546-3551). The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular ssDNA. This results in multiple substitutions of the first template into the second. *See*, Levichkin (1995) *Mol. Biology* 29:572-577; Jung (1992) *Gene* 121:17-24.

In some embodiments of the invention, shuffled nucleic acids obtained by use of the recursive recombination methods of the invention, are put into a cell and/or organism for screening. Shuffled genes can be introduced into, for example, bacterial cells, yeast cells, fungal cells vertebrate cells, invertebrate cells or plant cells for initial screening. *Bacillus* species (such as *B. subtilis* and *E. coli* are two examples of suitable bacterial cells into which one can insert and express shuffled genes which provide for convenient shuttling to other cell types (a variety of vectors for shuttling material between these bacterial cells and eukaryotic cells are available; *see*, Sambrook, Ausubel and Berger, *all supra*). The shuffled genes can

be introduced into bacterial, fungal or yeast cells either by integration into the chromosomal DNA or as plasmids.

Bacterial, plant, animal and yeast systems are preferred in the present invention. For example, in one embodiment, shuffled genes can be introduced into plant or animal cells for production purposes (it will be appreciated that transgenic plants are, increasingly, an important source of industrial enzymes), or can be introduced into a plant or animal cell for therapeutic purposes. Thus, a transgene of interest can be modified using the recursive sequence recombination methods of the invention *in vitro* and reinserted into the cell for *in vivo/in situ* selection for the new or improved property, in bacteria, eukaryotic cells, or whole eukaryotic organisms.

In Vivo DNA Shuffling Formats

In some embodiments of the invention, DNA substrate molecules, e.g., those comprising codon modifications relative to a wild-type sequence, are introduced into cells, where the cellular machinery directs their recombination. For example, a library of mutants is constructed and screened or selected for mutants with improved phenotypes by any of the techniques described herein.

The DNA substrate molecules encoding the best candidates are recovered by any of the techniques described herein, then fragmented and used to transfect a plant host and screened or selected for improved function. If further improvement is desired, the DNA substrate molecules are recovered from the host cell, such as by PCR, and the process is repeated until a desired level of improvement is obtained. In some embodiments, the fragments are denatured and reannealed prior to transfection, coated with recombination stimulating proteins such as recA, or co-transfected with a selectable marker such as Neo^R to allow the positive selection for cells receiving recombined versions of the gene of interest. Methods for *in vivo* shuffling are described in, for example, PCT application WO 98/13487 and WO 97/20078. The efficiency of *in vivo* shuffling can be enhanced by increasing the copy number of a gene of interest in the host cells.

Whole Genome Shuffling

In one embodiment, the selection methods herein are utilized in a "whole genome shuffling" format. An extensive guide to the many forms of whole genome shuffling is found in the pioneering application to the inventors and their co-workers entitled

"Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," PCT/US99/15972, by del Cardayre et al. Any codon-altered set of nucleic acids can be used to transform cells, which can then be shuffled by in a whole genome format.

In brief, whole genome shuffling makes no presuppositions at all regarding what nucleic acids may confer a desired property. Instead, entire genomes (e.g., from a genomic -library, or isolated from an organism) are shuffled in cells and selection protocols applied to the cells. These genomes can be spiked with any desired set of nucleic acids, including codon-modified nucleic acids.

Assays

The relevant assay for selection of a desired property of a codon-modified nucleic acid will depend on the application. Many assays which detect activity for proteins, receptors, ligands, cells and the like are known. Formats include binding to immobilized components, cell or organismal viability, production of reporter compositions, and the like.

In the high throughput assays of the invention, it is possible to screen up to several thousand different shuffled variants in a single day. In particular, each well of a microtiter plate can be used to run a separate assay, or, if concentration or incubation time effects are to be observed, every 5-10 wells can test a single variant. Thus, a single standard microtiter plate can assay about 100 (e.g., 96) reactions. If 1536 well plates are used, then a single plate can easily assay from about 100- about 1500 different reactions. It is possible to assay several different plates per day; assay screens for up to about 6,000-20,000 different assays (i.e., involving different nucleic acids, encoded proteins, concentrations, etc.) is possible using the integrated systems of the invention. More recently, microfluidic approaches to reagent manipulation have been developed, e.g., by Caliper Technologies (Mountain View, CA).

In one aspect, library members, e.g., cells, viral plaques, spores or the like, are separated on solid media to produce individual colonies (or plaques). Using an automated colony picker (e.g., the Q-bot, Genetix, U.K.), colonies or plaques are identified, picked, and up to 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or viruses in plaque applications). The time the pin is in the colony, the number of dips to

inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each can be controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. The glass balls in the microtiter plates act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter. Clones from cultures of interest can be cloned by limiting dilution. As also described supra, plaques or cells constituting libraries can also be screened directly for production of proteins, either by detecting hybridization, protein activity, protein binding to antibodies, or the like.

The ability to detect a subtle increase in the performance of a shuffled library member over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size, a prescreen that increases the number of mutants processed by, e.g., 10-fold can be used. The goal of the primary screen is to quickly identify mutants having equal or better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture for subsequent analysis.

A number of well known robotic systems have also been developed for solution phase chemistries useful in assay systems. These systems include automated workstations like the automated synthesis apparatus developed by Takeda Chemical Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Hewlett-Packard, Palo Alto, Calif.) which mimic the manual synthetic operations performed by a scientist. Any of the above devices are suitable for use with the present invention, e.g., for high-throughput screening of molecules encoded by codon-altered nucleic acids. The nature and implementation of modifications to these devices (if any) so that they can operate as discussed herein with reference to the integrated system will be apparent to persons skilled in the relevant art.

High throughput screening systems are commercially available (*see, e.g.,* Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures including all sample and reagent pipetting, liquid

dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization.

The manufacturers of such systems provide detailed protocols the various high throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like. Microfluidic approaches to reagent manipulation have also been developed, e.g., by Caliper Technologies (Mountain View, CA).

Optical images viewed (and, optionally, recorded) by a camera or other recording device (*e.g.*, a photodiode and data storage device) are optionally further processed in any of the embodiments herein, *e.g.*, by digitizing the image and/or storing and analyzing the image on a computer. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, *e.g.*, using PC (Intel x86 or pentium chip- compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (*e.g.*, SUN™ work station) computers.

One conventional system carries light from the assay device to a cooled charge-coupled device (CCD) camera, in common use in the art. A CCD camera includes an array of picture elements (pixels). The light from the specimen is imaged on the CCD. Particular pixels corresponding to regions of the specimen (*e.g.*, individual hybridization sites on an array of biological polymers) are sampled to obtain light intensity readings for each position. Multiple pixels are processed in parallel to increase speed. The apparatus and methods of the invention are easily used for viewing any sample, e.g., by fluorescent or dark field microscopic techniques.

Software elements for manipulating strings of characters which correspond to codon-modified nucleic acids can be used to direct synthesis of oligonucleotides relevant to shuffling of codon-modified nucleic acids. Integrated systems comprising these and other useful features, e.g., a digital computer with additional features such as high-throughput liquid control software, image analysis software, data interpretation software, a robotic liquid control armature for transferring solutions from a source to a destination operably linked to the digital computer, an input device (e.g., a computer keyboard) for entering data to the

digital computer to control high throughput liquid transfer by the robotic liquid control armature an image scanner for digitizing label signals from labeled assay components, or the like are a feature of the invention.

In one aspect, the invention provides an integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database. As discussed theroughout, there are a variety of sequence database programs for aligning and manipulating sequences. In addition, standard text manipulation software such as word processing software (e.g., Microsft Word™ or Corel Wodperfect™) and database software (e.g., spreadsheet software such as Microsoft Excel™, Corel Quattro Pro™, or database programs such as Microsoft Access™ or Paradox™) can be used in conjuction with a user interface (e.g., a GUI in a standard operating system such as a Windows, Macintosh or LINUX system) to manipulate strings of characters. Specialized alignment programs such as BLAST can also be incorporated into the systems of the invention for alignment of codon-altered nucleic acids (or corresponding character strings).

In addition to the integrated system elements mentioned above, the integrated system can also include an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium. Typically, the synthesizer is programmed to synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to test shuffled codon-modified DNAs, including in an iterative process.

An assay, kit or system utilizing a use of any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits will optionally additionally comprise instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally comprise one or more of the following: (1) a shuffled codon-modified component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein; (3) one or more assay component; (4) a container for holding nucleic acids or enzymes, other nucleic acids, transgneic plants, animals, cells, or the like, (5) packaging materials and (6) software fixerd in a computer readable medium comprising sequences corresponding to one or more codon-altered nucleic acid character string.

In a further aspect, the present invention provides for the use of any component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and materials described above can be used in various combinations. All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

WHAT IS CLAIMED IS:

1. A method of making codon altered nucleic acids, the method comprising:
(i) providing a first nucleic acid sequence, which nucleic acid sequence encodes a first polypeptide sequence;
(ii) providing a plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide or a modified form thereof; and,
(iii) recombining the plurality of codon-altered nucleic acid sequences to produce a target codon altered nucleic acid, which target codon altered nucleic acid encodes a second protein.

2. The method of claim 1, wherein at least one of the plurality of codon altered nucleic acid sequences does not hybridize to the first nucleic acid under stringent hybridization conditions.

3. The method of claim 1, further comprising shuffling a nucleic acid comprising a subsequence consisting of the first nucleic acid, or a substantially identical variant thereof, with one or more of the plurality of codon altered nucleic acids, or with the target codon altered nucleic acid.

4. The method of claim 1, the method further comprising the step of:
(iv) screening the second protein for a structural or functional property.

5. The method of claim 1, the method further comprising the steps of:
(iv) screening the second protein for a structural or functional property, and,
(v) comparing the structural or functional property of the second protein to a structural or functional property of the first protein.

6. The method of claim 1, wherein the second polypeptide has a structural or functional property equivalent or superior to the first polypeptide.

7. The method of claim 1, wherein the first and second polypeptide are homologous.

8. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered nucleic acids.

9. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered conservatively modified nucleic acids.

10. A library of codon altered conservatively modified nucleic acids produced by the method of claim 9.

11. The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered non-conservatively modified nucleic acids.

12. A library of codon altered conservatively modified nucleic acids produced by the method of claim 11.

13. The method of claim 1, wherein the plurality of codon altered nucleic acids is derived from a plurality of forms of the first nucleic acid.

14. The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise at least three codon altered nucleic acids.

15. The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise one or more of the following structural features:

(a) codon usage divergence for each of the codon altered nucleic acids of 50% or more as compared to the first nucleic acid;

(b) codon usage divergence for each of the codon altered nucleic acids of 75% or more as compared to the first nucleic acid;

(c) codon usage divergence for each of the codon altered nucleic acids of 90% or more as compared to the first nucleic acid;

(d) maximal codon usage divergence for each of the codon altered nucleic acids as compared to the first nucleic acid;

(e) non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid;

(f) a lack of high stringency hybridization between one or more of the codon altered nucleic acid and the first nucleic acid; and,

(g) modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide.

16. The method of claim 1, wherein the percent identity between the second protein and the first protein is lower than the percent identity between two of the plurality of codon altered nucleic acids.

17. The method of claim 1, wherein the first nucleic acid encodes a protein selected from: EPO, G-CSF, a viral envelope protein, a cytokine, and a phosphatase.

18. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are isolated nucleic acids.

19. The target codon altered nucleic acid produced by the method of claim 1.

20. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are nucleic acids present in cells.

21. The cells produced by the method of claim 20.

22. The method of claim 1, wherein each of the codon altered nucleic acid sequences comprises at least two nucleotide differences when compared to the first nucleic acid.

23. The method of claim 1, further comprising introducing the target codon altered nucleic acid into a cell, or into a vector or virus.

24. The cell, vector or virus produced by the method of claim 23.

25. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus.

26. The attenuated virus produced by the method of claim 25.

27. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus, which attenuated virus produces an immune response upon infection by the virus in a mammal.

28. The attenuated virus produced by the method of claim 27.

29. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a retroviral genome to produce an attenuated retrovirus, which attenuated retrovirus produces an immune response upon infection by the retrovirus in a mammal.

30. The attenuated retrovirus produced by the method of claim 29.

31. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector.

32. The viral vector produced by the method of claim 31.

33. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector, which vector requires *trans* complementation for replication, and which vector has a reduced rate of reversion to a replicative form as compared to a corresponding viral vector which lacks a subsequence corresponding to the target codon altered nucleic acid.

34. The viral vector produced by the method of claim 33.

35. The method of claim 33, wherein the vector comprises viral elements from one or more of: a lentivirus, an adenovirus, a herpes virus, and an adeno-associated virus.

36. A method of making a library of codon-altered nucleic acids, the method comprising:

(i) selecting a first nucleic acid sequence, which nucleic acid sequence encodes a first polypeptide sequence; and,

(ii) making a plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide or a modified form thereof, wherein the plurality of codon altered nucleic acids comprise the library.

37. A codon-altered library made by the method of claim 36.

38. The library of claim 37, wherein said library comprises at least 2 codon altered nucleic acids.

39. The library of claim 37, wherein said library comprises at least 5 codon altered nucleic acids.

40. The library of claim 37, wherein said library comprises at least 10 codon altered nucleic acids.

41. The library of claim 37, wherein said library comprises at least 100 codon altered nucleic acids.

42. A kit comprising the library of claim 25 and one or more of: a container and instructional materials providing method step instructions for recombining two or more of members of the library.

43. The method of claim 41, further comprising recombining said plurality of codon altered nucleic acids to produce a shuffled codon-altered library.

44. The codon altered library made by the method of claim 43.

45. The method of claim 36, wherein said nucleic acids encode a protein selected from EPO, a cytokine, a phosphatase, and a viral envelope protein.

46. A composition comprising a plurality of codon altered nucleic acids, each of which encode a first polypeptide or a modified form thereof.

47. A library of codon-altered nucleic acids, comprising a plurality of codon-altered nucleic acids derived from a plurality of homologous nucleic acids.

48. The library of claim 47, wherein said plurality of codon altered nucleic acids recombine in vitro at an increased rate compared to said plurality of homologous nucleic acids.

49. The library of claim 47, wherein the level of identity among said plurality of codon-altered nucleic acids is at least as high as among a plurality of polypeptides encoded by said plurality of homologous nucleic acids.

50. An integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database.

51. The integrated system of claim 50, further comprising an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium, which synthesizer is programmed to synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

Translation of Monkey EPO cDNA

Sau3A

```
GATCCCGCGCCCCCTGGACAGCCGCCCTCTCCTCCAGGCCCGTGGGGCTGGCCCTGCCC

CGCTGAACTTCCCGGGATGAGGACTCCCGGTGTGGTCACCGCGCGCCTAGGTCGCTGAG

                                -27                         -20
                                Met Gly  Val  His  Glu  Cys  Pro  Ala  Trp
GGACCCCGGCCAGGCGCGGAGATG GGG  GTG  CAC  GAA  TGT  CCT  GCC  TGG
                                      -10
 Leu Trp Leu Leu Leu Ser Leu Val Ser Leu Pro Leu Gly Leu Pro
 CTG TGG CTT CTC CTG TCT CTC GTG TCG CTC CCT CTG GGC CTC CCA
          -1  +1                                    10
 Val Pro Gly Ala Pro Pro Arg Leu Ile Cys Asp Ser Arg Val Leu
 GTC CCG GGC GCC CCA CCA CGC CTC ATC TGT GAC AGC CGA GTC CTG
                                  20
 Glu Arg Tyr Leu Leu Glu Ala Lys Glu Ala Glu Asn Val Thr Met
 GAG AGG TAC CTC TTG GAG GCC AAG GAG GCC GAG AAT GTC ACG ATG
          30                                        40
 Gly Cys Ser Glu Ser Cys Ser Leu Asn Glu Asn Ile Thr Val Pro
 GGC TGT TCC GAA AGC TGC AGC TTG AAT GAG AAT ATC ACC GTC CCA
```

Fig. 1

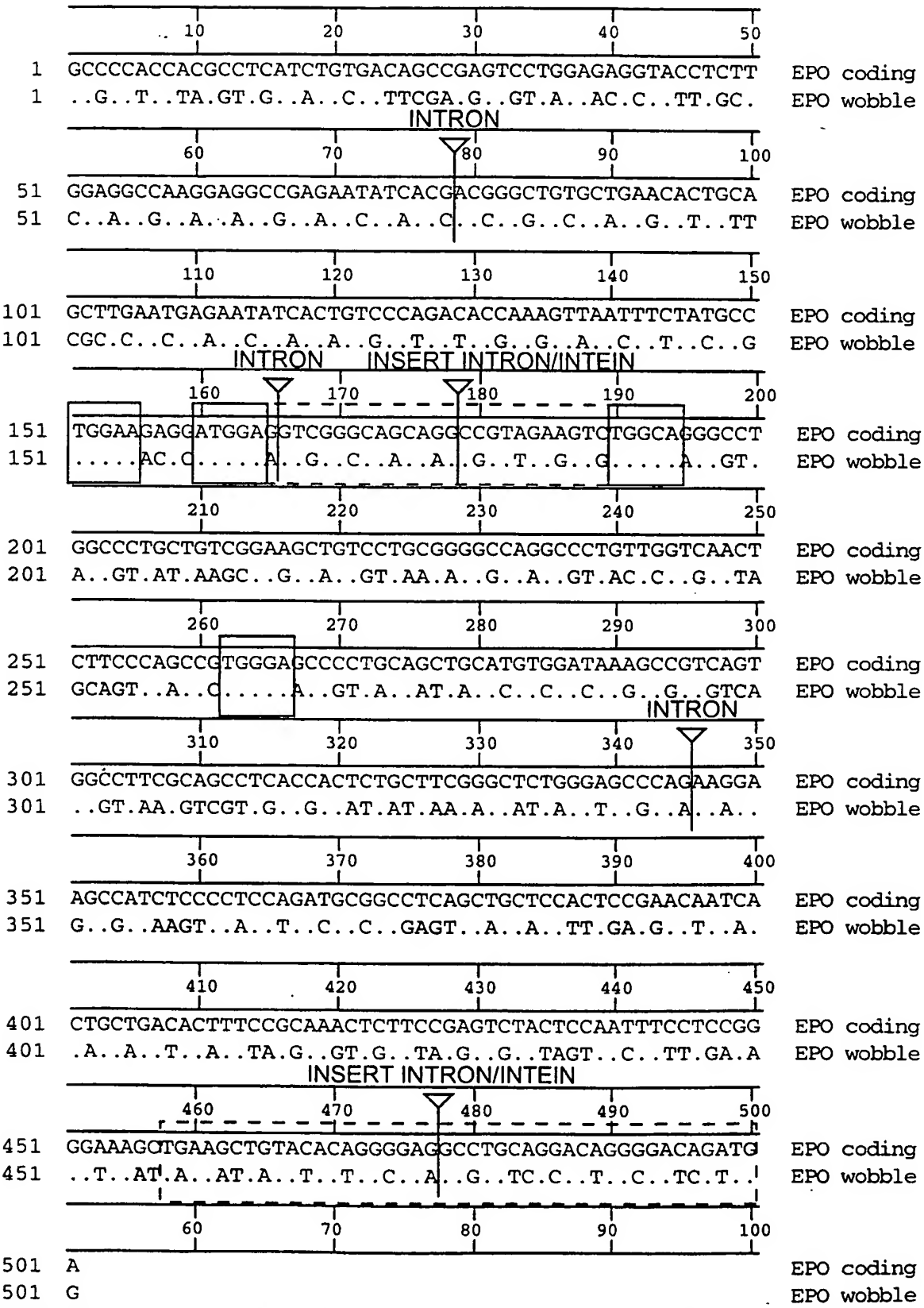
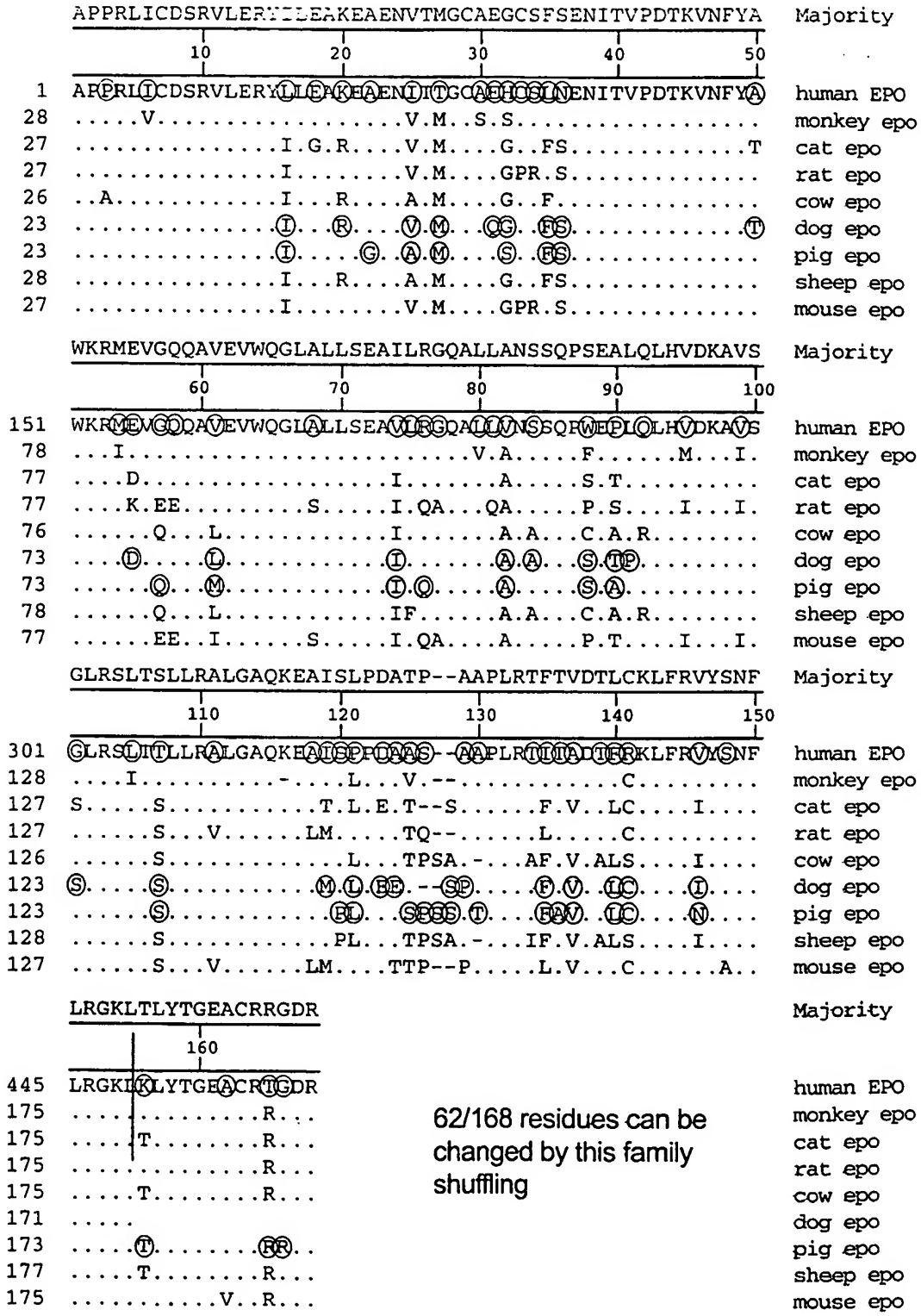

10 20 30 40 50
1 GCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGAGAGGTACCTCTT EPO coding
1 ..G..T..TA.GT.G..A..C..TTCGA.G..GT.A..AC.C..TT.GC. EPO wobble
INTRON
60 70 80 90 100
51 GGAGGCCAAGGAGGCCGAGAATATCACGACGGGCTGTGCTGAACACTGCA EPO coding
51 C..A..G..A..A..G..A..C..A..C..C..G..C..A..G..T..TT EPO wobble
110 120 130 140 150
101 GCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTTAATTTCTATGCC EPO coding
101 CGC.C..C..A..C..A..A..G..T..T..G..G..A..C..T..C..G EPO wobble
INTRON INSERT INTRON/INTEIN
160 170 180 190 200
151 TGGAAGAGGATGGAGGTCGGGCAGCAGGCCGTAGAAGTCTGGCAGGGCCT EPO coding
151 .....AC.C.....A..G..C..A..A..G..T..G..G.....A..GT. EPO wobble
210 220 230 240 250
201 GGCCCTGCTGTCGGAAGCTGTCCTGCGGGGCCAGGCCCTGTTGGTCAACT EPO coding
201 A..GT.AT.AAGC..G..A..GT.AA.A..G..A..GT.AC.C..G..TA EPO wobble
260 270 280 290 300
251 CTTCCCAGCCGTGGGAGCCCCTGCAGCTGCATGTGGATAAAGCCGTCAGT EPO coding
251 GCAGT..A..C.....A..GT.A..AT.A..C..C..C..G..G..GTCA EPO wobble
INTRON
310 320 330 340 350
301 GGCCTTCGCAGCCTCACCACTCTGCTTCGGGCTCTGGGAGCCCAGAAGGA EPO coding
301 ..GT.AA.GTCGT.G..G..AT.AT.AA.A..AT.A..T..G..A..A.. EPO wobble
360 370 380 390 400
351 AGCCATCTCCCCTCCAGATGCGGCCTCAGCTGCTCCACTCCGAACAATCA EPO coding
351 G..G..AAGT..A..T..C..C..GAGT..A..A..TT.GA.G..T..A. EPO wobble
410 420 430 440 450
401 CTGCTGACACTTTCCGCAAACTCTTCCGAGTCTACTCCAATTTCCTCCGG EPO coding
401 .A..A..T..A..TA.G..GT.G..TA.G..G..TAGT..C..TT.GA.A EPO wobble
INSERT INTRON/INTEIN
460 470 480 490 500
451 GGAAAGCTGAAGCTGTACACAGGGGAGGCCTGCAGGACAGGGGACAGATG EPO coding
451 ..T..AT.A..AT.A..T..T..C..A..G..TC.C..T..C..TC.T.. EPO wobble
60 70 80 90 100
501 A EPO coding
501 G EPO wobble
Decoration 'Decoration #1': Hide (as '.') residues that match EPO coding exactly

Fig. 2



Decoration 'Decoration #1': Hide (as '.') residues that match EPO coding exactly

Regions of Relative Conservation

Fig. 3

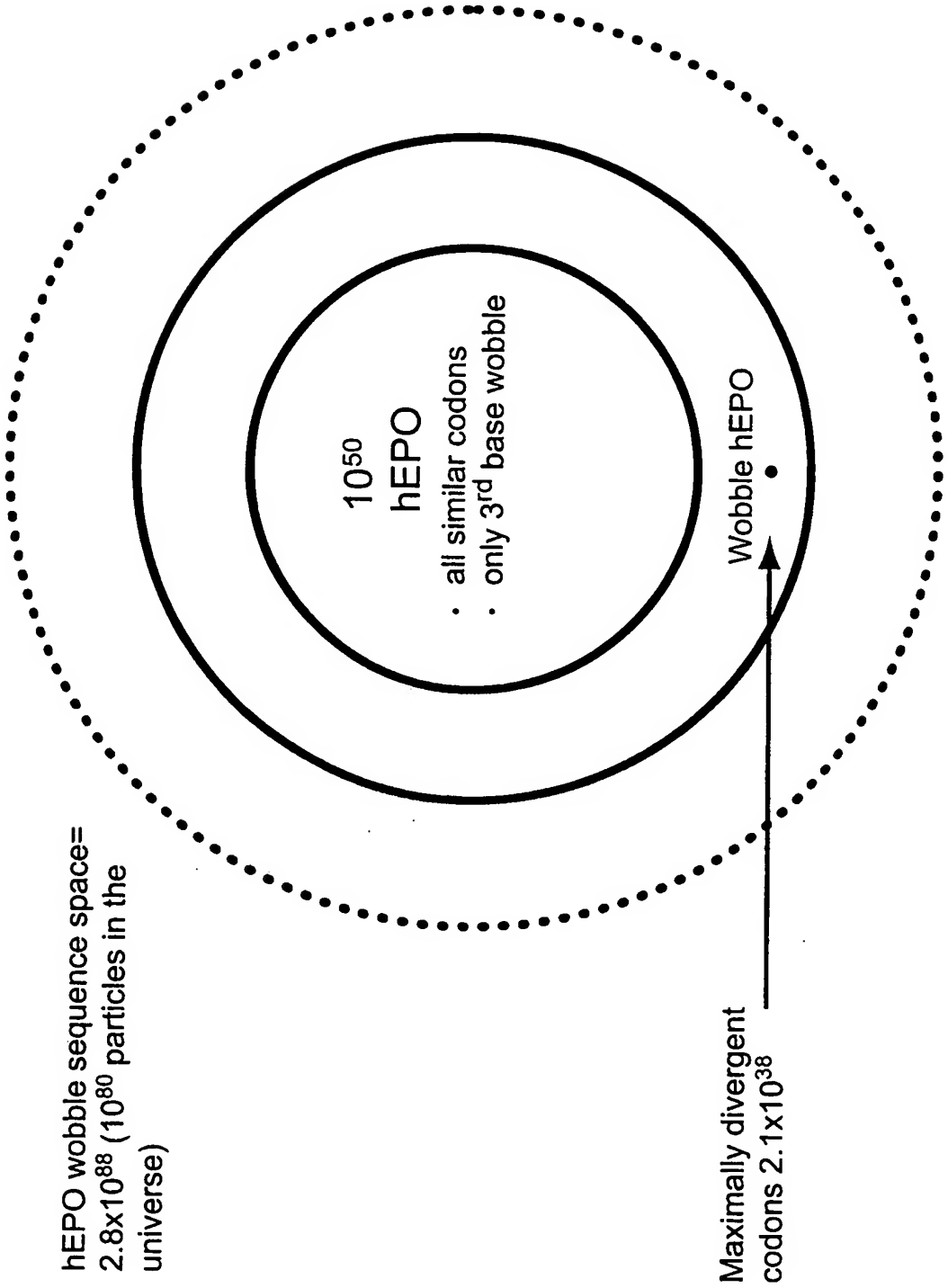

hEPO wobble sequence space= 2.8x10^88 (10^80 particles in the universe)
10^50 hEPO
· all similar codons
· only 3rd base wobble
Wobble hEPO
Maximally divergent codons 2.1x10^38

Fig. 4

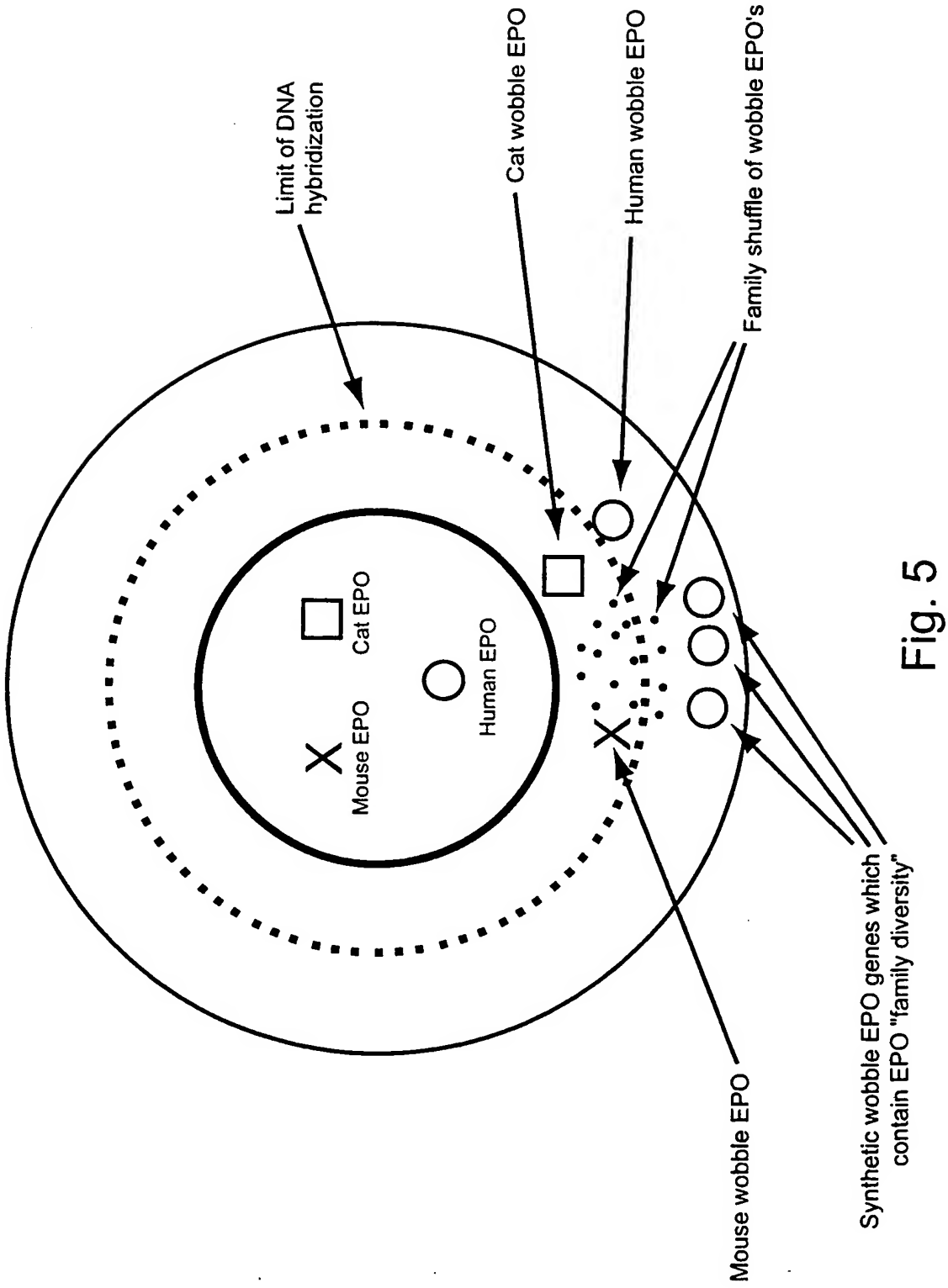
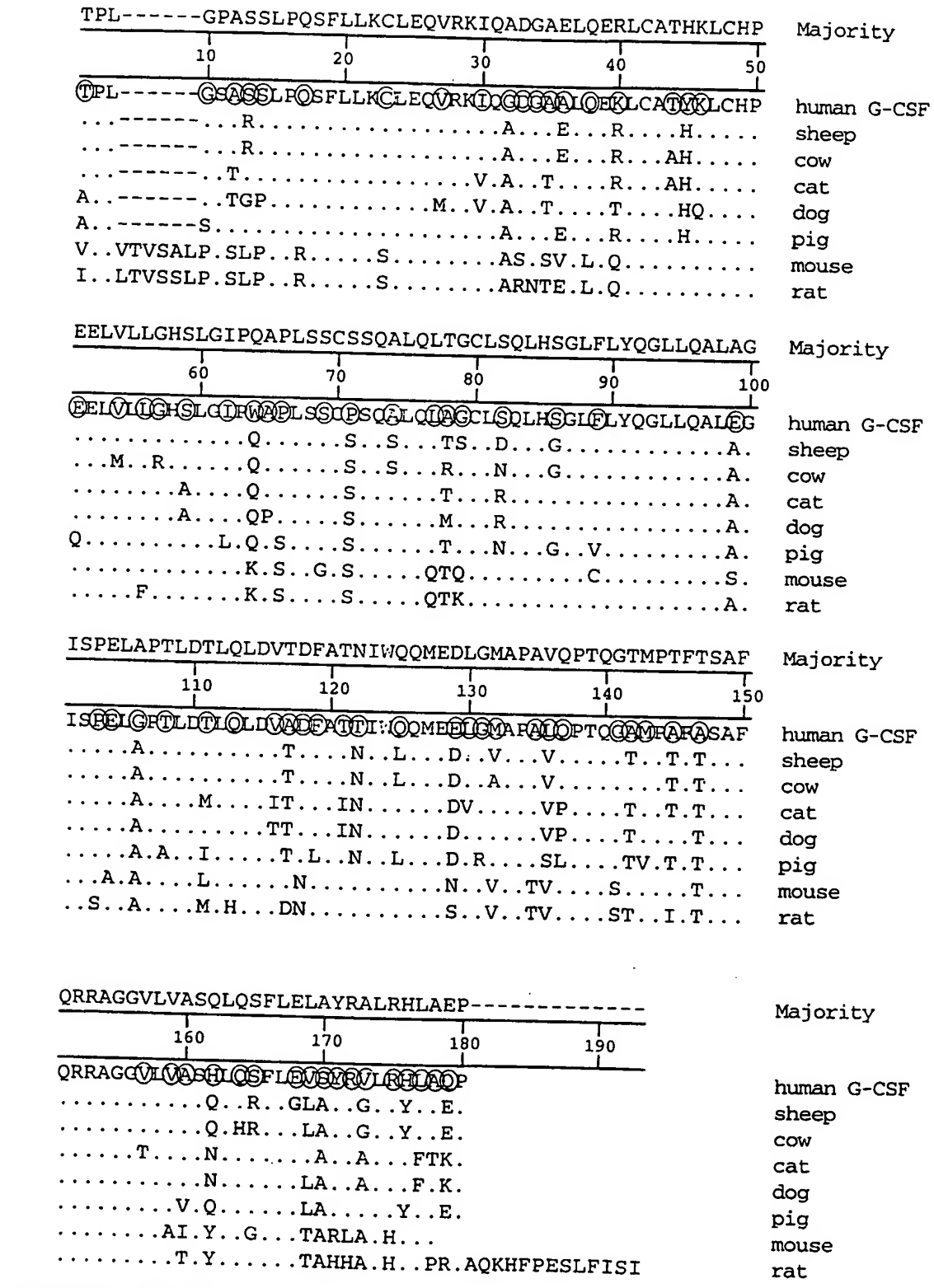


Fig. 5



Decoration 'Decoration #1': Hide (as '.') residues that match EPO coding exactly

Fig. 6

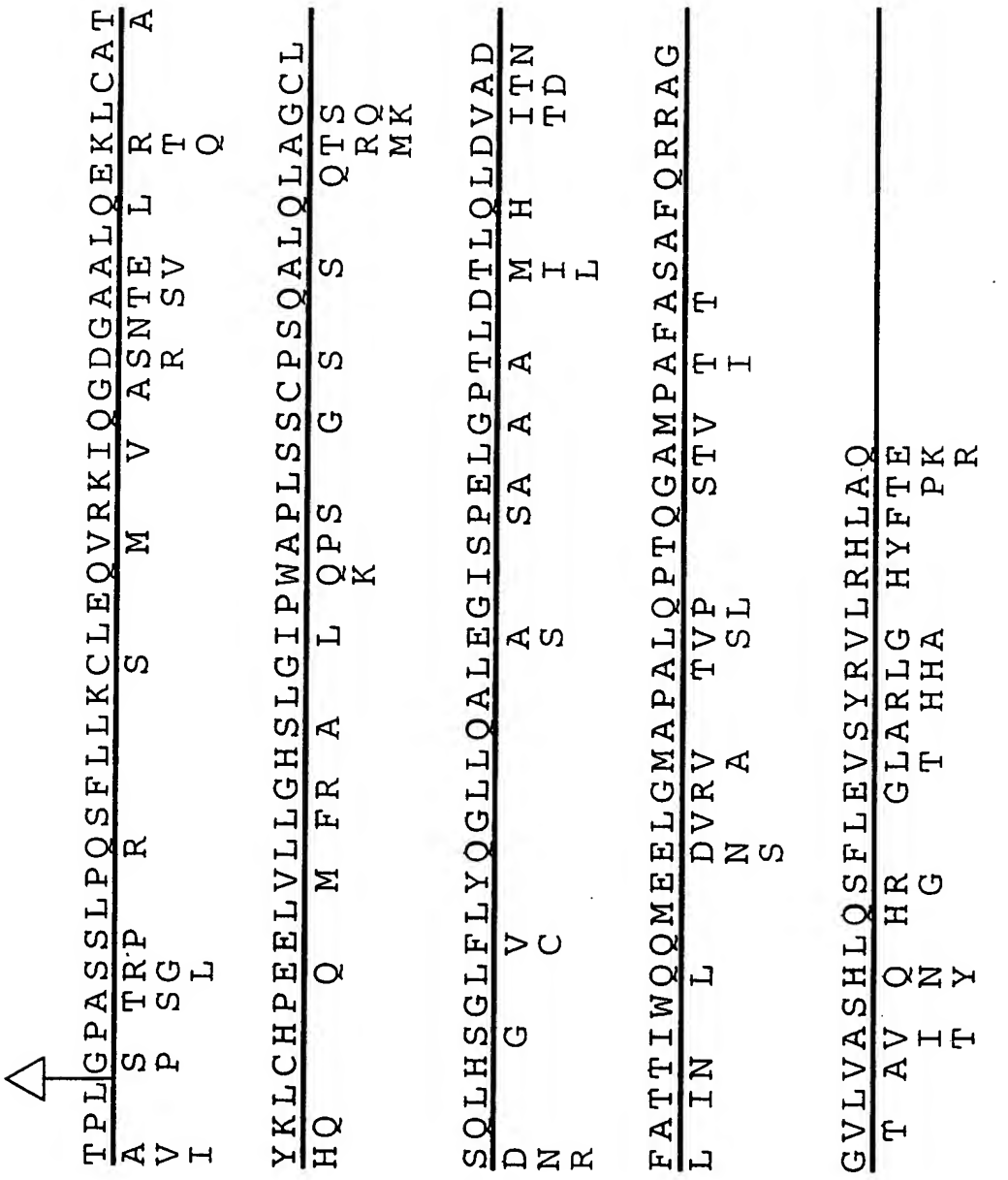
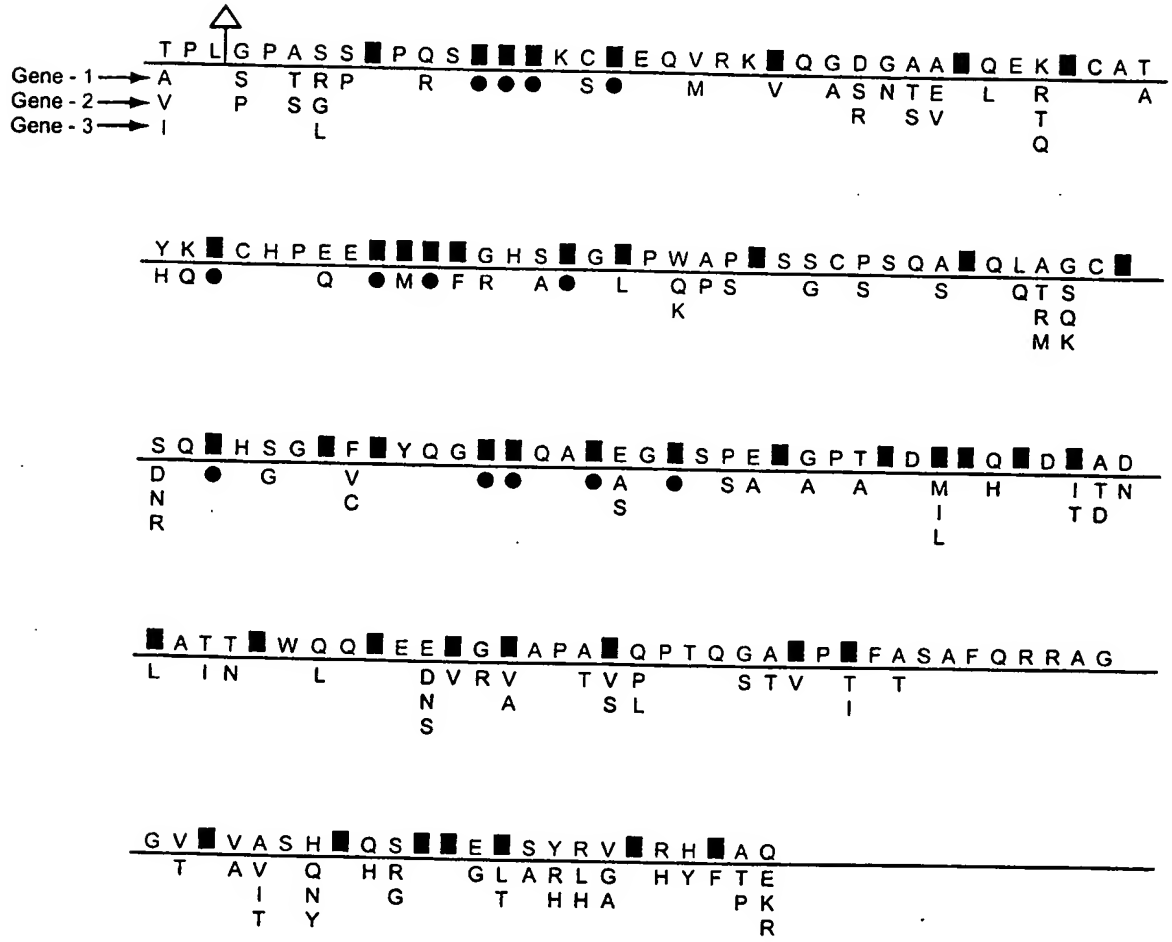


Fig. 7

G-CSF Family Diversity



● = hydrophobic core residues in conserved stretches of the protein

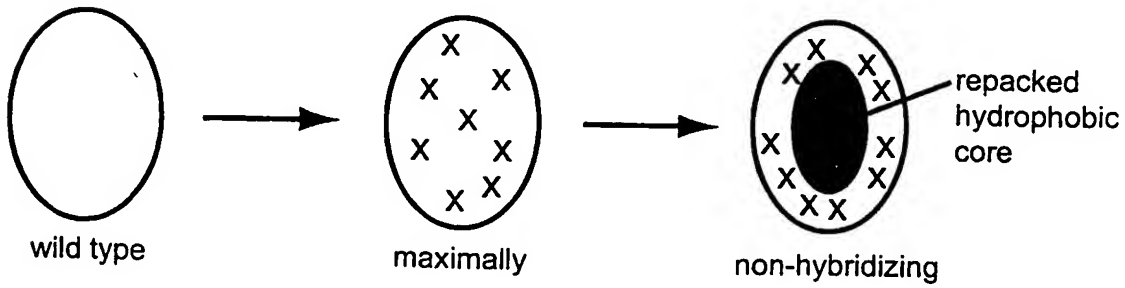


Fig. 8

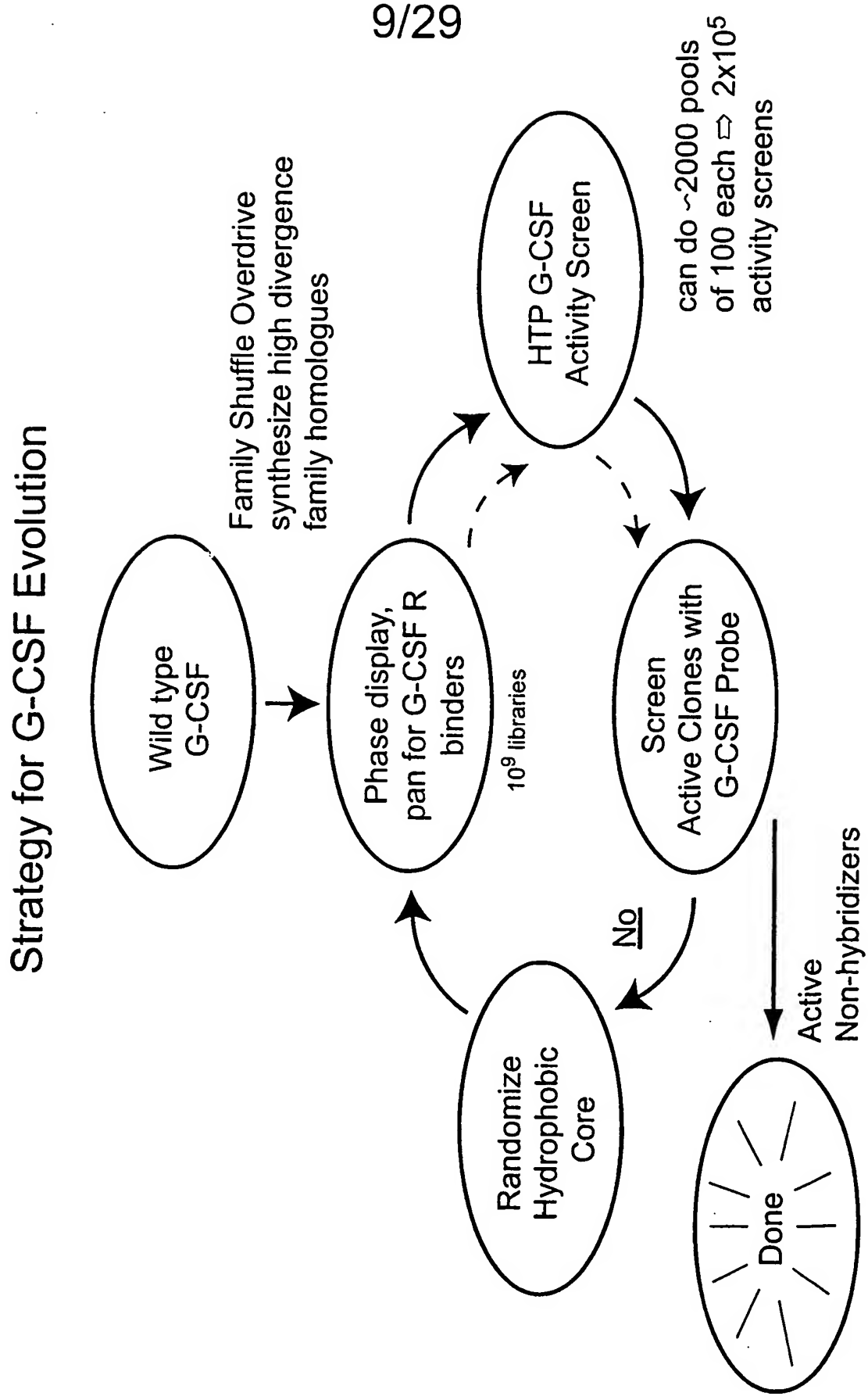


Fig. 9

Assembling of Alkaline Phosphatase

```
F1lamb   AACCCTCCAGTTCCGAACCCCATATGATGATCACCCTGCGTAAACTGCCG
F1ompa   AACCCTCCAGTTCCGAACCCCATATGAAAAAAACCGCT
F1ompt   AACCCTCCAGTTCCGAACCCATATACATATGCGTGCTAAA
F1pelb   AACCCTCCAGTTCCGAACCCCATATGAAATACCTGCTGCCGACC
F1phoa   AACCCTCCAGTTCCGAACCCGATATACATATGAAACAGTC
F2lamb   TGGTGTTATGTCTGCTCAGG CDATGGCDGTDGAYTTYCAY CTGGTTCCGGTTGAAGAGGA
F2ompa   GGCTGGTTTCGCTACCGTTGCDCARGCDGCDCCDAARGAYCTGGTTCCGGTTGAAGAGGA
F2ompt   CACCCCGATCGCTATCTCTTCYTTYGCDTCYACYGGYTCYCTGGTTCCGGTTGAAGAGGA
F2pelb   GCTGCTGGCTGCTCAGCCGGCDATGGCDATGGAYATYGGYCTGGTTCCGGTTGAAGAGGA
F2phoa   TGCCGCTGCTGTTCACCCCGGTDACYAARGCDGCDCARGTDCTGGTTCCGGTTGAAGAGGA
F3       CCCGGCTTTCTGGAACCGTC ARGCDGCDCARGCDCTGGAY GTTGCTAAAAAACTGCAGCC
F4       ACGTTATCCTGTTCCTGGGT GAYGGYATGGGYGTDCCDAC CGTTACCGCTACCCGTATCC
F5       AAACTGGGTCCGGAAACCCC DCTGGCDATGGAYCARTTYC CGTACGTTGCTCTGTCTAAA
F6       GGTTCCGGACTCTGCTGGTA CYGCDACYGCDTAYCTGTGY GGTGTTAAAGGTAACTACCG
F7       CTGCTCGTTACAACCAGTGC AARACYACYCGYGGYAAYGA AGTTACCTCTGTTATGAACC
F8       TCTGTTGGTGTTGTTACCAC YACYCGYGTDCARCAYGCDT CTCCGGCTGGTGCTTACGCT
F9       GTACTCTGACGCTGACCTGC CDGCDGAYGCDCARATGAAY GGTTGCCAGGACATCGCTGC
F10      ACATCGACGTTATCCTGGGT GGYGGYCGYAARTAYATGTT CCCGGTTGGTACCCCGGACC
F11      TCTGTTAACGGTGTTCGTAA RCGYAARCARAAYCTGGTDC AGGCTTGGCAGGCTAAACAC
F12      GAACCGTACCGCTCTGCTGC ARGCDGCDGAYGAYTCYTCY GTTACCCACCTGATGGGTCT
F13      AATACAACGTTCAGCAGGAC CAYACYAARGAYCCDACYCT GCAGGAAATGACCGAAGTTG
F14      AACCCGCGTGGTTTCTACCT GTTYGTDGARGGYGGYCGYA TCGACCACGGTCACCACGAC
F15      GACCGAAGCTGGTATGTTCG AYAAYGCDATYGCDAARGCD AACGAACTGACCTCTGAACT
F16      CCGCTGACCACTCTCACGTT TTYTCYTTYGGYGGYTAYAC CCTGCGTGGTACCTCTATCT
F17      GCTCTGGACTCTAAATCTTA YACYTCYATYCTGTAYGGYA ACGGTCCGGGTTACGCTCTG
F18      CGTTAACGACTCTACCTCTG ARGAYCCDTCYTAYCARCAR CAGGCTGCTGTTCCGCAGGC
F19      AAGACGTTGCTGTTTTCGCT CGYGGYCCDCARGCDCAYCT GGTTCACGGTGTTGAAGAAG
F20      ATGGCTTTCGCTGGTTGCGT DGARCCDTAYACYGAYTGYA ACCTGCCGGCTCCGACCACC
F21      TGCTCACCTGGCTGCTTMAC CDCCDCCDCTGGCDCTGCTG GCTGGTGCTATGCTGCTCCTC

R1       TTCCGCCTCTAGAGAATTCT TARTACAGRGTHGGHGCCAG GAGGAGCAGCATAGCACCAGCC
R2       AAGCAGCCAGGTGAGCAG CGTCHGGRATRGARGTHGCR GTGGTCGGAGCCGGCAGGTT
R3       CGCAACCAGCGAAAGCCATG ATRTGHGCHACRAARGTYTC TTCTTCAACACCGTGAACCA
R4       GCGAAAACAGCAACGTCTTC RCCRCCRTGRGTYTCRGAHG CCTGCGGAACAGCAGCCTGC
R5       AGAGGTAGAGTCGTTAACGT CHGGRCGRGARCCRCCRCCC AGAGCGTAACCCGGACCGTT
R6       AAGATTTAGAGTCCAGAGCT TTRGAHGGHGCCAGRCCRAA GATAGAGGTACCACGCAGGG
R7       ACGTGAGAGTGGTCAGCGGT HACCAGRATCAGRGTRTCCA GTTCAGAGGTCAGTTCGTTA
R8       GAACATACCAGCTTCGGTCA GHGCCATRTAHGCYTTRTCR TCGTGGTGACCGTGGTCGAT
R9       GGTAGAAACCACGCGGGTTA CGRGAHACHACRCGCAGHGC AACTTCGGTCATTTCCTGCA
R10      TCCTGCTGAACGTTGTATTT CATRTCHGCHGGYTCRAACA GACCCATCAGGTGGGTAACA
R11      CAGCAGAGCGGTACGGTTCC AHACRTAYTGHGCRCCYTGR TGTTTAGCCTGCCAAGCCTG
R12      TACGAACACCGTTAACAGAA GCRTCRTCHGGRTAYTCHGG GTCCGGGGTACCAACCGGGA
R13      CCCAGGATAACGTCGATGTC CATRTTRTTHACCAGYTGHG CAGCGATGTCCTGGCAACCG
R14      CAGGTCAGCGTCAGAGTACC ARTTRCGRTTHACRGTRTGH GCGTAAGCACCAGCCGGAGA
R15      TGGTAACAACACCAACAGAT TTRCCHGCYTTYTTHGCRCG GTTCATAACAGAGGTAACTT
R16      CACTGGTTGTAACGAGCAGC HGCRGAHACRCCRATRGTRC GGTAGTTACCTTTAACACCG
R17      ACCAGCAGAGTCCGGAACCT GRCGRTCHACRTTRTARGTY TTAGACAGAGCAACGTACGG
R18      GGGTTTCCGGACCCAGTTTA CCRTTCATYTGRCCYTTCAG GATACGGGTAGCGGTAACGG
R19      CCCAGGAACAGGATAACGTT YTTHGCHGCRGTYTGRATHG GCTGCAGTTTTTTAGCAACG
R20      ACGGTTCCAGAAAGCCGGGT CTTCTTCTTCAACCGGAACC AG
R21lamb   CCTGAGCAGACATAACACCA GCHGCHACHGCHACHGCCAG CGGCAGTTTACGCAGGGTGA
R21phoa   ACCGGGGTGAACAGCAGCGGCAGCAGHGCCAGHGCRATRGTRGACTGTTTCATATGTATATC
R21pelb   GCCGGCTGAGCAGCCAGCAGCAGCAGRCCHGCHGCHGCGGTCGGCAGCAGGTAGTTCA
R21ompt   AAGAGATAGCGATCGGGGTGGTCAGHACRATRCCCAGCAGTTTAGCACGCATATGTATAT
R21ompa   CAACGGTAGCGAAACCAGCCAGHGCHACHGCRATHGCRATAGCGGTTTTTTTCATATGTA
```

circularization:

```
5'AGAATTCTCTAGAGGCGGAA  ACTCTCCAACTCCCAGGTT 5' F1 all
                        TGAGAGGTTGAGGGTCCAA  TTGGGAGGTCAAGGCTTGGG 5'
```

Fig. 10

Fig. 11

Vaccination with evolution-defective viruses

Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration

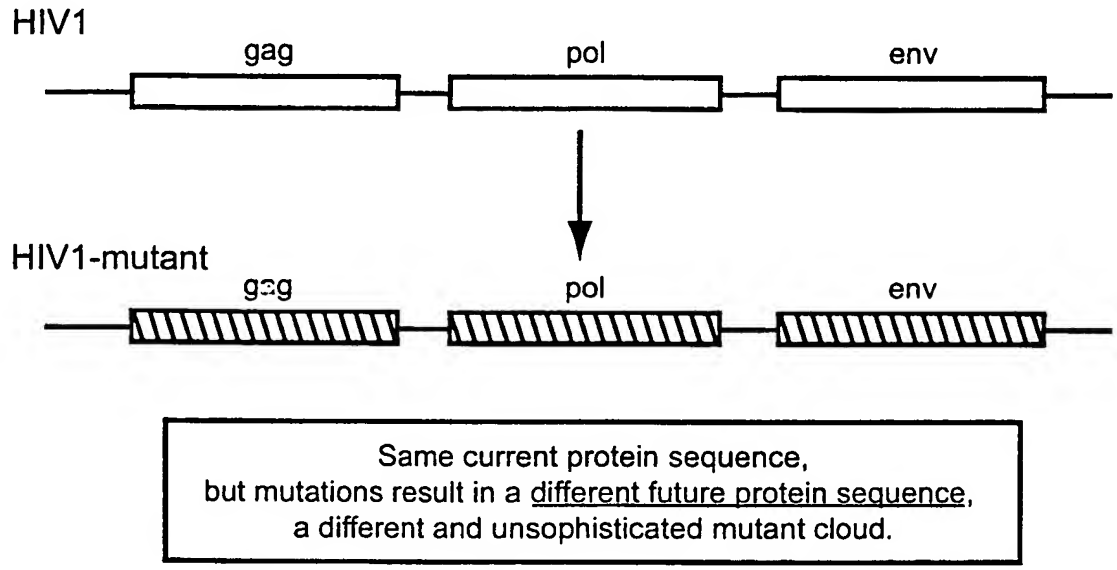
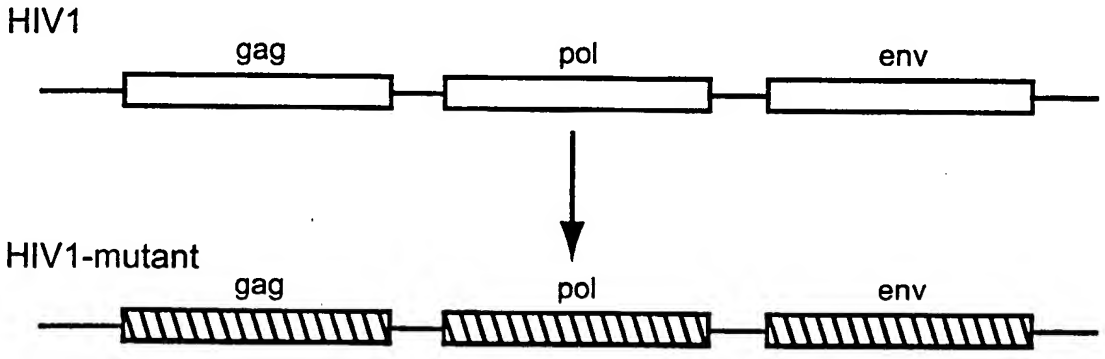


Fig. 12

Vaccination with evolution-defective viruses

Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration



Same current protein sequence,
but mutations result in a different future protein sequence,
a different and unsophisticated mutant cloud.

Maximizing the difference requires the following changes:

Most Effective

Ser: TCX ⟷ AGY
Arg: CGX ⟷ AGR
Leu: CTX ⟷ TTR

Moderately Effective

Thr, Pro, Val, Ala, Gly:

--R ⟷ --Y

Ile:

ATY ⟷ ATA

Least Effective

Phe, Tyr, Cys, His, Asn, Asp:

--C ⟷ --T

Lys, Gln, Glu:

--A ⟷ --G

Unchangeable

Trp: TGG
Met: ATG

Fig. 14

Arg

AGA	CGT	AGA	CGC	AGA	CGA	AGA	CGG
TGA .	AGT S	TGA .	AGC S	TGA .	AGA S	TGA .	AGG S
CGA R	GGT G	CGA R	GGC G	CGA R	GGA G	CGA R	GGG G
GGA G	TGT C	GGA G	TGC C	GGA G	TGA .	TGG W	CGG R
AAA K	CAT H	AAA K	CAC H	AAA K	CAA Q	CAG Q	CGG P
ACA T	CCT P	ACA T	CCC P	ACA T	CCA P	CCG P	CGG R
ATA I	CTT L	ATA I	CTC L	ATA I	CTA L	CTG L	CGG R
AGC S	CGA R	AGC S	CGA R	AGC S	CGC R	CGA R	CGC R
AGT S	CGC R	AGT S	CGT R	AGT S	CGT R	CGT R	CGC R
AGG R	CGG R	AGG R	CGG R	AGG R	CGG R	CGC R	CGC R

AGG	CGT	AGG	CGC	AGG	CGA	AGG	CGG
TGG W	AGT S	TGG W	AGC S	TGG W	AGA S	TGG W	AGG S
CGG R	GGT G	CGG R	GGC G	CGG R	GGA G	CGG R	GGG G
GGG G	TGT C	GGG G	TGC C	GGG G	TGA .	GGG G	TGG W
AAG K	CAT H	AAG K	CAC H	AAG K	CAA Q	AAG K	CAG Q
ACG T	CCT P	ACG T	CCC P	ACG T	CCA P	ACG T	CCG P
ATG M	CTT L	ATG M	CTC L	ATG M	CTA L	ATG M	CTG L
AGA R	CGA R	AGA R	CGA R	AGA R	CGC R	AGA R	CGA R
AGC S	CGC R	AGC S	CGT R	AGC S	CGT R	AGC S	CGT R
AGT S	CGG R	AGT S	CGG R	AGT S	CGG R	AGT S	CGC R

Fig. 16A

Ser

AGI		ICI		ICC		AGC		ICT		ICG		AGC	
CGT	R	ACT	T	ACC	T	CGC	R	ACA	T	ACG	T	CGC	R
GGT	G	CCT	P	CCC	P	GGC	G	CCA	P	CCG	P	GGC	G
TGT	C	GCT	A	GCC	A	TGC	C	GCA	A	GCG	A	TGC	C
AAT	N	TAT	Y	TAC	Y	AAC	N	TAA	.	TAG	.	AAC	N
ACT	T	TGT	C	TGC	C	ACC	T	TGA	.	TGG	W	ACC	T
ATT	I	TTT	F	TTC	F	ATC	I	TTA	L	TTG	L	ATC	I
AGA	R	TCA	S	TCA	S	AGA	R	TCC	S	TCA	S	AGA	R
AGG	R	TCG	S	TCG	S	AGG	R	TCG	S	TCC	S	AGG	R
AGC	S	TCC	S	TCC	S	AGT	S	TCT	S	TCT	S	AGT	S

AGI		ICI		ICC		AGC		ICT		ICG		AGC	
CGT	R	ACT	T	ACC	T	CGC	R	ACA	T	ACG	T	CGC	R
GGT	G	CCT	P	CCC	P	GGC	G	CCA	P	CCG	P	GGC	G
TGT	C	GCT	A	GCC	A	TGC	C	GCA	A	GCG	A	TGC	C
AAT	N	TAT	Y	TAC	Y	AAC	N	TAA	.	TAG	.	AAC	N
ACT	T	TGT	C	TGC	C	ACC	T	TGA	.	TGG	W	ACC	T
ATT	I	TTT	F	TTC	F	ATC	I	TTA	L	TTG	L	ATC	I
AGA	R	TCA	S	TCA	S	AGA	R	TCC	S	TCA	S	AGA	R
AGG	R	TCG	S	TCG	S	AGG	R	TCG	S	TCC	S	AGG	R
AGC	S	TCC	S	TCC	S	AGT	S	TCT	S	TCT	S	AGT	S

Fig. 16B

Leu

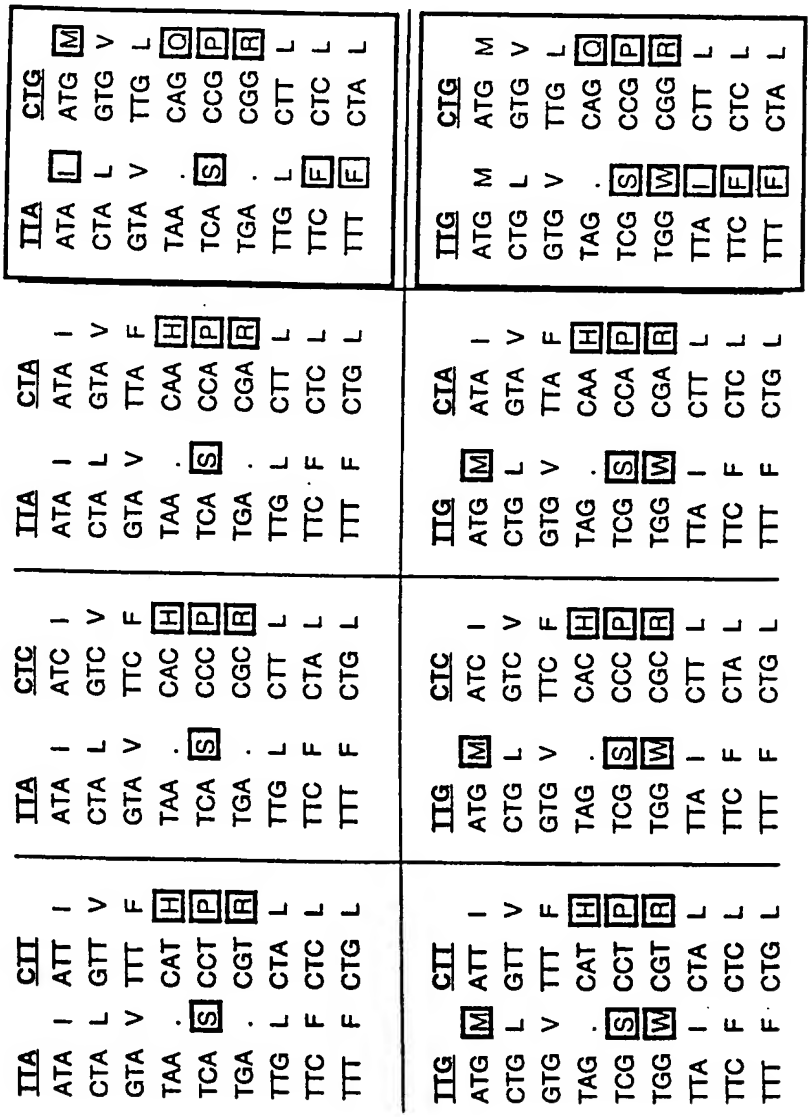
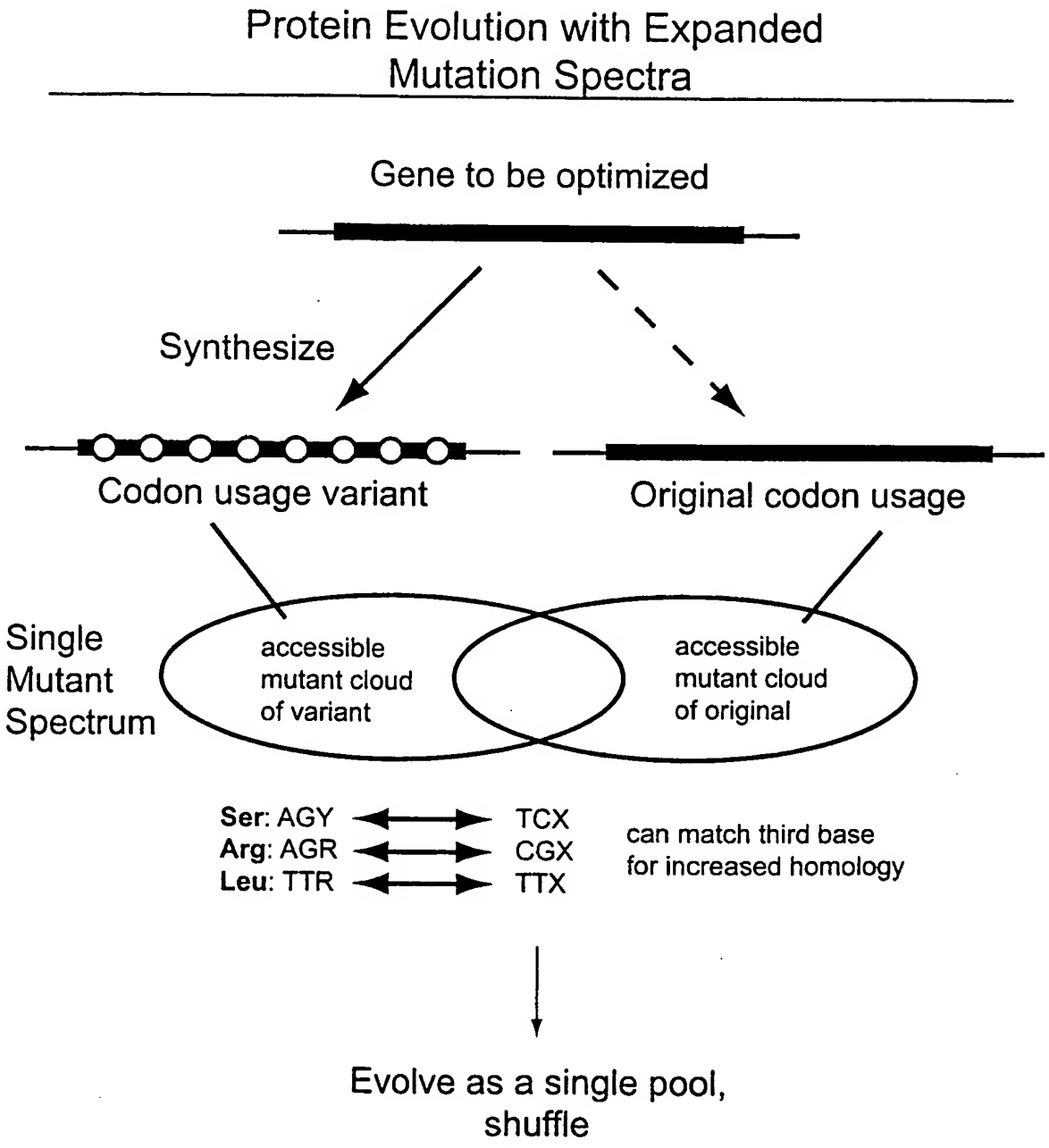

IIA CTI
ATA I ATT I
CTA L GTT V
GTA V TTT F
TAA . CAT H
TCA S CCT P
TGA . CGT R
TTG L CTA L
TTC F CTC L
TTT F CTG L
IIA CTC
ATA I ATC I
CTA L GTC V
GTA V TTC F
TAA . CAC H
TCA S CCC P
TGA . CGC R
TTG L CTT L
TTC F CTA L
TTT F CTG L
IIA CTA
ATA I ATA I
CTA L GTA V
GTA V TTA F
TAA . CAA H
TCA S CCA P
TGA . CGA R
TTG L CTT L
TTC F CTC L
TTT F CTG L
IIA CTG
ATA I ATG M
CTA L GTG V
GTA V TTG L
TAA . CAG Q
TCA S CCG P
TGA . CGG R
TTG L CTT L
TTC F CTC L
TTT F CTA L
IIG CTI
ATG M ATT I
CTG L GTT V
GTG V TTT F
TAG . CAT H
TCG S CCT P
TGG W CGT R
TTA I CTA L
TTC F CTC L
TTT F CTG L
IIG CTC
ATG M ATC I
CTG L GTC V
GTG V TTC F
TAG . CAC H
TCG S CCC P
TGG W CGC R
TTA I CTT L
TTC F CTA L
TTT F CTG L
IIG CTA
ATG M ATA I
CTG L GTA V
GTG V TTA F
TAG . CAA H
TCG S CCA P
TGG W CGA R
TTA I CTT L
TTC F CTC L
TTT F CTG L
IIG CTG
ATG M ATG M
CTG L GTG V
GTG V TTG L
TAG . CAG Q
TCG S CCG P
TGG W CGG R
TTA I CTT L
TTC F CTC L
TTT F CTA L

Fig. 16C



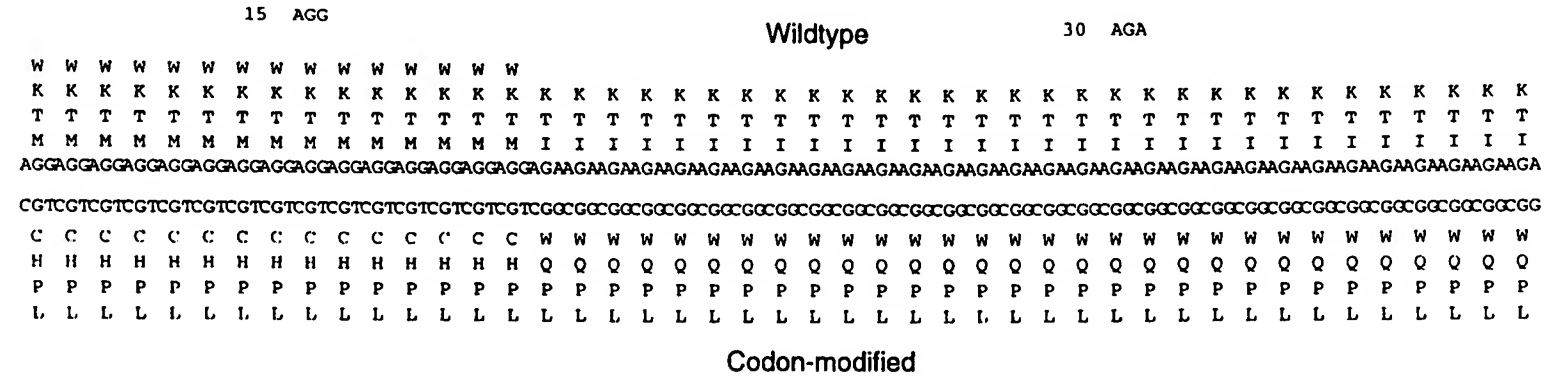
Best synthesis method: all possible codons based on codon synthesis

Fig. 17

Altering the evolution potential of sequences by codon modification

HIV envelope protein (857 amino acids)

Potential mutant resulting from wildtype sequence versus codon modified sequence in arg, leu and serine codons:



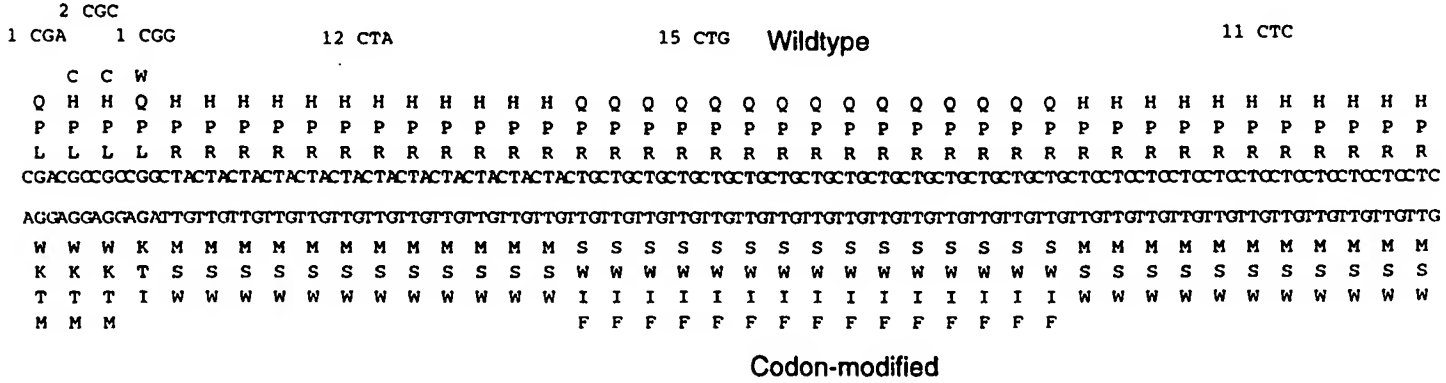


Fig. 18A

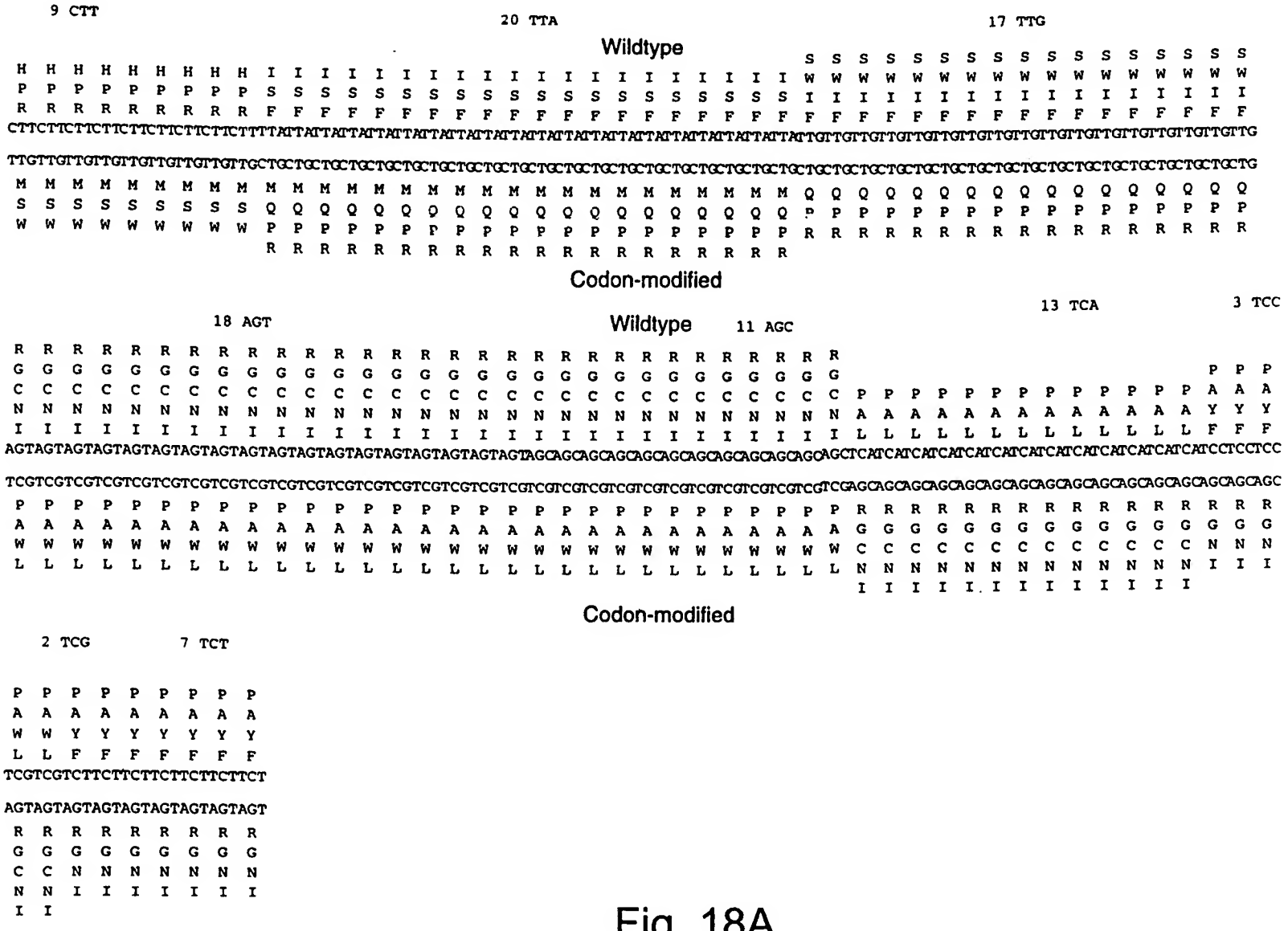

9 CTT
20 TTA
17 TTG
Wildtype
Codon-modified
18 AGT
Wildtype
11 AGC
13 TCA
3 TCC
Codon-modified
2 TCG
7 TCT

Fig. 18A

22/29

FINAL ALIGNMENT NEW ENV VS OLD ENV

Saturday, August 17, 1996 2:03 PM
Lipman-Pearson Protein Alignment
kTuple: 2; Gap Penalty: 4; Gap Length Penalty: 12

Page 1

Seq1(1>509) new env pro	Seq2(1>509) old env protein	Similarity Index	Gap Number	Gap Length	Consensus Length
(1>509)	(1>509)	100.0	0	0	509

```
          ↓10        ↓20        ↓30        ↓40        ↓50        ↓60        ↓70
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
          ↳10        ↳20        ↳30        ↳40        ↳50        ↳60        ↳70
          ↓80        ↓90        ↓100       ↓110       ↓120       ↓130       ↓140
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
          ↳80        ↳90        ↳100       ↳110       ↳120       ↳130       ↳140
          ↓150       ↓160       ↓170       ↓180       ↓190       ↓200       ↓210
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
          ↳150       ↳160       ↳170       ↳180       ↳190       ↳200       ↳210
          ↓220       ↓230       ↓240       ↓250       ↓260       ↓270       ↓280
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
          ↳220       ↳230       ↳240       ↳250       ↳260       ↳270       ↳280
          ↓290       ↓300       ↓310       ↓320       ↓330       ↓340       ↓350
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
          ↳290       ↳300       ↳310       ↳320       ↳330       ↳340       ↳350
          ↓360       ↓370       ↓380       ↓390       ↓400       ↓410       ↓420
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
          ↳360       ↳370       ↳380       ↳390       ↳400       ↳410       ↳420
          ↓430       ↓440       ↓450       ↓460       ↓470       ↓480       ↓490
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
          ↳430       ↳440       ↳450       ↳460       ↳470       ↳480       ↳490
          ↓500
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
          ↳500
```

Fig. 18B

DNA HOMOLOGY

Saturday, August 17, 1996 4:39 PM
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1527) new env	Seq2(1>1527) old env	Similarity Index	Gap Number	Gap Length	Consensus Length
(10>1454)	(10>1454)	61.7	0	0	1445

```
 10        20        30        40        50        60
GTBACBGTBTTTTACGGAGTTCCTGCTTGGAAAAACGCAACBATMCCDTTRTTTTGCGCT
GT AC GT TT TA GG GT CC GC TGGAA AA GC AC AT CC  T TT TG GC
GTAACAGTATTCTATGGTGTACCAGCATGGAAGAATGCGACAATTCCCCTCTTCTGTGCA
 10        20        30        40        50        60
 70        80        90        100       110       120
ACAAAAAACCGHGATACVTGGGGBACBACBCAGTGTCTCCCGACAACGACGACTATAGY
AC AA AA  G GA AC TGGGG AC AC CA TG  T CC GA AA GA GA TA
ACCAAGAATAGGGACACTTGGGGAACAACACAATGCTTGCCAGATAATGATGATTACTCA
 70        80        90        100       110       120
 130       140       150       160       170       180
GAGCTHGCBATAAACGTAACTGAAGCATTCGACGCVTGGGACAACACBGTDACTGAGCAG
GA  T GC AT AA GT AC GA GC TT GA GC TGGGA AA AC GT AC GA CA
GAATTGGCAATCAATGTCACAGAGGCTTTTGATGCTTGGGATAATACAGTCACAGAACAA
 130       140       150       160       170       180
 190       200       210       220       230       240
GCTATTGAAGACGTHTGGAATTTRTTCGAGACTAGTATAAAACCATGCGTTAAGTTRACD
GC AT GA GA GT TGGAA  T TT GA AC    AT AA CC TG GT AA  T AC
GCAATAGAGGATGTGTGGAACCTCTTTGAAACATCCATTAAGCCCTGTGTAAAACTCACC
 190       200       210       220       230       240
 250       260       270       280       290       300
CCBTTGTGCATTGCTATGAGGTGCAACAAGACVGAAACBGACCGHTGGGGVCTCACTGGT
CC  T TG AT GC ATGAG TG AA AA AC GA AC GA  G TGGGG  T AC GG
CCACTATGTATAGCAATGAGATGTAATAAAACTGAGACAGATAGGTGGGGTTTGACAGGA
 250       260       270       280       290       300
 310       320       330       340       350       360
AATGCTGGTACBACBACBACBGCBATYACBACTACCGCTACBCCTTCVGTTGCBGAGAAC
AA GC GG AC AC AC AC GC AT AC AC AC GC AC CC    GT GC GA AA
AACGCAGGGACAACAACAACAGCAATAACAACAACAGCAACACCAAGTGTAGCAGAAAAT
 310       320       330       340       350       360
 370       380       390       400       410       420
GTYATYAACGAGTCAAACCCATGTATTAAGAACAACTCVTGCGCBGGDCTYGAGCAAGAA
GT AT AA GA    AA CC TG AT AA AA AA    TG GC GG  T GA CA GA
GTTATAAATGAAAGTAATCCTTGCATAAAAAATAATAGTTGTGCAGGCTTGGAACAGGAG
 370       380       390       400       410       420
 430       440       450       460       470       480
CCDATGATTGGVTGCAAGTTCAATATGACBGGCCTTAACCGCGATAAGAAAAAGGAGTAC
CC ATGAT GG TG AA TT AA ATGAC GG  T AA  G GA AA AA AA GA TA
CCCATGATAGGTTGTAAATTTAACATGACAGGGTTAAATAGGGACAAAAAAAAAGAATAT
 430       440       450       460       470       480
 490       500       510       520       530       540
AACGAGACBTGGTACAGTCGTGACCTTATATGCGATCAAAGYGCHAACGAGTCAGAATCC
AA GA AC TGGTA    G GA  T AT TG GA CA    GC AA GA    GA
AATGAAACATGGTATTCAAGAGATTTAATCTGTGAGCAGTCAGCGAATGAAAGTGAGAGT
 490       500       510       520       530       540
 550       560       570       580       590       600
AAGTGCTATATGCACCACTGCAATACDTCVGTAATACAGGAGAGTTGCGATAAACACTAC
AA TG TA ATGCA CA TG AA AC    GT AT CA GA    TG GA AA CA TA
AAATGTTACATGCATCATTGTAACACCAGTGTTATTCAAGAATCCTGTGACAAGCATTAT
 550       560       570       580       590       600
```

Fig. 18C

Saturday, August 17, 1996 4:39 PM Page 2
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1>1527) new env DNA	Seq2(1>1527) old env.pro	Similarity Index	Gap Number	Gap Length	Consensus Length
(10>1454)	(10>1454)	61.7	0	0	1445

```
f610       f620       f630       f640       f650       f660
TGGGACGCVATACGTTTCCGCTATTGCGCTCCHCCBGGVTACGCVCTHTTRCGCTGCAAC
TGGGA GC AT  G TT  G TA TG GC CC CC GG TA GC  T  T  G TG AA
TGGGATGCTATTAGATTTAGATACTGTGCACCGCCAGGTTATGCTTTGCTTAGGTGTAAT
t610       t620       t630       t640       t650       t660
f670       f680       f690       f700       f710       f720
GACAGTAACTACCTBGGDTTCGCVCCVAATTGCAGYAAAGTTGTCGTAAGCAGYTGTACB
GA    AA TA  T GG TT GC CC AA TG    AA GT GT GT       TG AC
GATTCAAATTATTTAGGCTTTGCTCCTAACTGTTCTAAGGTAGTGGTTTCTTCATGCACA
t670       t680       t690       t700       t710       t720
f730       f740       f750       f760       f770       f780
CGBATGATGGAAACCCAGACGAGCACATGGTTCGGDTTTAACGGVACVCGHGCTGAGAAC
 G ATGATGGA AC CA AC    AC TGGTT GG TT AA GG AC  G GC GA AA
AGAATGATGGAGACGCAAACCTCTACTTGGTTTGGCTTCAATGGTACTAGGGCAGAAAAT
t730       t740       t750       t760       t770       t780
f790       f800       f810       f820       f830       f840
CGTACTTATATMTACTGGCACGGDAAGTCVAACCGTACGATTATATCGCTCAACAAATAC
 G AC TA AT TA TGGCA GG AA    AA  G AC AT AT     T AA AA TA
AGAACATACATTTATTGGCATGGCAAAAGTAATAGAACCATAATTAGCTTGAATAAGTAT
t790       t800       t810       t820       t830       t840
f850       f860       f870       f880       f890       f900
TACAACTTYACTATGCGTTGCCGTCGCCCTGGBAACAAAACBGTVCTBCCBGTAACGATA
TA AA  T AC ATG G TG  G  G CC GG AA AA AC GT  T CC GT AC AT
TATAATCTAACAATGAGATGTAGAAGACCAGGAAATAAGACAGTTTTACCAGTCACCATT
t850       t860       t870       t880       t890       t900
f910       f920       f930       f940       f950       f960
ATGACTGGCCTHGTDTTTCACAGYCAACCDATTAACGAACGTCCTAAGCAAGCDTGGTGT
ATG C GG  T GT TT CA    CA CC AT AA GA  G CC AA CA GC TGGTG
ATGTCAGGGTTGGTCTTCCATTCGCAGCCCATAAATGAGAGACCAAAACAGGCCTGGTGC
t910       t920       t930       t940       t950       t960
f970       f980       f990       f1000      f1010      f1020
TGGTTCGAGGGTTCGTGGAAGAAAGCGATACAAGAGGTHAAAGAGACDCTHGTGAAGCAC
TGGTT GA GG    TGGAA AA GC AT CA GA GT AA GA AC  T GT AA CA
TGGTTTGAAGGAAGCTGGAAAAAGGCCATCCAGGAAGTGAAGGAAACCTTGGTCAAACAT
t970       t980       t990       t1000      t1010      t1020
f1030      f1040      f1050      f1060      f1070      f1080
CCGCGCTACACHGGBACVAACGACACVCGHAAGATAAACTTGACTGCACCTGCBGGBGGB
CC  G TA AC GG AC AA GA AC  G AA AT AA  T AC GC CC GC GG GG
CCCAGGTATACGGGAACTAATGATACTAGGAAAATTAATCTAACAGCTCCAGCAGGAGGA
t1030      t1040      t1050      t1060      t1070      t1080
f1090      f1100      f1110      f1120      f1130      f1140
GACCCBGAGGTGACATTCATGTGGACTAACTGCAGRGGBGAGTTTCTBTACTGTAAGATG
GA CC GA GT AC TT ATGTGGAC AA TG  G GG GA TT  T TA TG AA ATG
GATCCAGAAGTCACTTTTATGTGGACAAATTGTCGAGGAGAATTCTTATATTGCAAAATG
t1090      t1100      t1110      t1120      t1130      t1140
f1150      f1160      f1170      f1180      f1190      f1200
AACTGGTTCTTRAACTGGGTBGAAGATCGBGATCAGAAAGGAGGGCGTTGGAAGCAGCAG
AA TGGTT  T AA TGGGT GA GA  G GA CA AA GG GG  G TGGAA CA CA
AATTGGTTTCTTAATTGGGTAGAGGACAGAGACCAAAAGGGTGGCAGATGCAAACAACAA
t1150      t1160      t1170      t1180      t1190      t1200
```

Fig. 18C

Saturday, August 17, 1996 4:39 PM Page 3
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

Seq1(1>1527) new env DNA	Seq2(1>1527) old env.pro	Similarity Index	Gap Number	Gap Length	Consensus Length
(10>1454)	(10>1454)	61.7	0	0	1445

```
 1210      1220      1230      1240      1250      1260
AACCGHAAGGAACAGCAAAAAAAGAACTACGTHCCBTGCCACATMCGBCAGATTATAAAT
AA  G AA GA CA CA AA AA AA TA GT CC TG CA AT  G CA AT AT AA
AATAGGAAAGAGCAACAGAAGAAAAATTATGTGCCATGTCATATTAGACAAATAATCAAC
 1210      1220      1230      1240      1250      1260
 1270      1280      1290      1300      1310      1320
ACCTGGCATAAGGTBGGDAAGAACGTBTACCTCCCACCCCGCGAGGGTGATTTRACBTGT
AC TGGCA AA GT GG AA AA GT TA  T CC CC  G GA GG GA  T AC TG
ACGTGGCACAAAGTAGGCAAAAATGTATATTTGCCTCCTAGGGAAGGAGACCTGACATGC
 1270      1280      1290      1300      1310      1320
 1330      1340      1350      1360      1370      1380
AACAGYACVGTTACATCATTGATTGCTGAAATYGACTGGATWAACTCDAACGAAACGAAC
AA    AC GT AC     T AT GC GA AT GA TGGAT AA    AA GA AC AA
AATTCCACTGTAACTAGTCTCATAGCAGAGATAGATTGGATCAATAGCAATGAGACCAAT
 1330      1340      1350      1360      1370      1380
 1390      1400      1410      1420      1430      1440
ATAACGATGTCVGCBGAAGTHGCBGAGTTYTACAGGCTCGAACTCGGTGACTATAAGCTB
AT AC ATG   GC GA GT GC GA  T TA  G  T GA  T GG GA TA AA  T
ATCACCATGAGTGCAGAGGTGGCAGAACTGTATCGATTGGAGTTGGGAGATTACAAATTA
 1390      1400      1410      1420      1430      1440
 1450
ATYGA
AT GA
ATAGA
 1450
```

Fig. 18C

26/29

MUTAGENIZED GENE IN CONTEXT

Saturday, August 17, 1996 2:06 PM
1.6 kb Env mutagenized
Page 1
Created: Friday, August 16, 1996 9:39 AM

```
          10        20        30        40
          |         |         |         |
aactccgaagaagactaaggctaatccatttcctgcatca 40
aacaagtaagtatgggatgtcttgggaatcagctgcttat 80
cgcgctcttgctagtaagtgttttagagatttgttgtGTv 120
CAgTAcGTbACbGTbTTtTAcGGaGTtCCTGCtTGGAAaA 160
AcGChACbATmCCdttrTTtTGcGCtACaAAaAAccghGA 200
        210       220       230       240
        |         |         |         |
tACvTGGGGbACbACbCAgTGtctcCCtGAcAAcGAcGAc 240
TAtagyGAgcthGCbATaAAcGTaACtGAaGCaTTcGAcG 280
CvTGGGAcAAcACbGTdACtGAgCAgGCtATtGAaGAcGT 320
hTGGAAtttrTTcGAgACtagtATaAAaCCaTGcGTtAAg 360
ttrACdCCbttgTGcATtGCtATGAGgTGcAAcAAgACvG 400
        410       420       430       440
        |         |         |         |
AaACbGAccghTGGGGvctcACtGGtAAtGCtGGtACbAC 440
bACbACbGCbATyACbACtACcGCtACbCCttcvGTtGCb 480
GAgAAcGTyATyAAcGAgtcaAAcCCaTGtATtAAgAAcA 520
ActcvTGcGCbGGdctyGAgCAaGAaCCdATGATtGGvTG 560
cAAgTTcAAtATGACbGGccttAAccgcGAtAAgAAaAAg 600
        610       620       630       640
        |         |         |         |
GAgTAcAAcGAgACbTGGTAcagtcgtGAccttATaTGcG 640
AaCAaagyGChAAcGAgtcaGAatccAAgTGcTAtATGCA 680
cCAcTGcAAtACdtcvGTaATaCAgGAgagtTGcGAtAAa 720
CAcTAcTGGGAcGCvATacgtTTccgcTAtTGcGCtCChC 760
CbGGvTAcGCvcthttrcgcTGcAAcGAcagtAAcTAcct 800
        810       820       830       840
        |         |         |         |
bGGdTTcGCvCCvAAtTGcagyAAaGTtGTcGTaagcagy 840
TGtACbcgbATGATGGAaACcCAgACgagcACaTGGTTcG 880
GdTTtAAcGGvACvcghGCtGAgAAccgtACtTAtATmTA 920
cTGGCAcGGdAAgtcvAAccgtACgATtATatcgctcAAc 960
AAaTAcTAcAActtgACtATGcgtTGccgtcgcCCtGGbA 1000
        1010      1020      1030      1040
        |         |         |         |
AcAAaACbGTvctbCCbGTaACgATaATGagtGGccthGT 1040
dTTtCAcagyCAaCCdATtAAcGAacgtCCtAAgCAaGCd 1080
TGGTGtTGGTTcGAgGGttcgTGGAAgAAaGCgATaCAaG 1120
AgGThAAaGAgACdcthGTgAAgCAcCCgcgcTAcAChGG 1160
bACvAAcGAcACvcghAAgATaAActtgACtGCaCCtGCb 1200
```

Fig. 18D

Saturday, August 17, 1996 2:06 PM
1.6 kb Env mutagenized


```
        1210       1220       1230       1240
GGbGGbGAcCCbGAgGTgACaTTcATGTGGACtAAcTGca 1240
grGGbGAgTTtctbTAcTGtAAgATGAAcTGGTTcttrAA 1280
cTGGGTbGAaGAtcgbGAtCAgAAaGGaGGgcgtTGGAAg 1320
CAgCAgAAccghAAgGAaCAgCAaAAaAAgAAcTAcGThC 1360
CbTGcCAcATmcgbCAgATtATaAAtACcTGGCAtAAgGT 1400
        1410       1420       1430       1440
bGGdAAgAAcGTbTAcctcCCaCCccgcGAgGGtGAtttr 1440
ACbTGtAAcagyACvGTtACatcattgATtGCtGAaATyG 1480
AcTGGATwAActcdAAcGAaACgAAcATaACgATGtcvGC 1520
bGAaGThGCbGAgttrTAcaggctcGAactcGGtGAcTAt 1560
AAgctbATyGAaATmACaCCtATaGGgctcGCgCCgACbt 1600
        1610       1620       1630       1640
cvGTbcgbcghTAtACdACtACaGGaGCgagtcgtAAcAA 1640
acgbggggtctttgtgctagggttcttgggttttctcgcg 1680
acagcaggttctgcaatgggcgcggcgtcGCtgacgctgt 1720
cggctcagtcccggactttgttggctgggat 1751
```

Fig. 18D

ENV Top oligos

1T CTAGTAAGTGTTTTAGAGATTTGTTGT

2T TTACGGAGTTCCTGCTTGGAAAAACGCHACBATMCCDTTRTTTTGCGCTACAAAAAACCG

3T AGTGTCTCCCTGACAACGACGACTATAGYGAGCTHGCBATAAACGTAACTGAAGCATTCG

4T ACTGAGCAGGCTATTGAAGACGTHTGGAATTTRTTCGAGACTAGTATAAAACCATGCGTT

5T TGCTATGAGGTGCAACAAGACVGAAACBGACCGHTGGGGVCTCACTGGTAATGCTGGTAC

6T CTACCGCTACTCCTTCAGTTGCBGAGAACGTYATYAACGAGTCAAACCCATGTATTAAGA

7T GAGCAAGAACCGATGATTGGVTGCAAGTTCAATATGACBGGCCTTAACCGCGATAAGAAA

8T GTACAGTCGTGACCTTATATGCGAACAAAGYGCHAACGAGTCAGAATCCAAGTGCTATAT

9T TAATACAGGAGAGTTGCGATAAACACTACTGGGACGCVATACGTTTCCGCTATTGCGCTC

10T CGCTGCAACGACAGTAACTACCTBGGDTTCGCVCCVAATTGCAGTAAAGTTGTCGTAAGC

11T AACCCAGACGAGCACATGGTTCGGDTTTAACGGVACVCGHGCTGAGAACCGTACTTATAT

12T ACCGTACGATTATATCGCTCAACAAATACTACAACTTGACTATGCGTTGCCGTCGCCCTG

13T GTAACGATAATGAGTGGCCTHGTDTTTCACAGYCAACCDATTAACGAACGTCCTAAGCAA

14T TTCGTGGAAGAAAGCGATACAAGAGGTHAAAGAGACDCTHGTGAAGCACCCGCGCTACAC

15T AGATAAACTTGACTGCACCTGCBGGBGGBGACCCBGAGGTGACATTCATGTGGACTAACT

16T TGTAAGATGAACTGGTTCTTRAACTGGGTBGAAGATCGBGATCAGAAAGGAGGGCGTTGG

17T ACAGCAAAAAAAGAACTACGTHCCBTGCCACATMCGBCAGATTATAAATACCTGGCATAA

18T TCCCACCCCGCGAGGGTGATTTRACBTGTAACAGYACVGTTACATCATTGATTGCTGAAA

19T GAAACGAACATAACGATGTCVGCBGAAGTHGCBGAGTTRTACAGGCTCGAACTCGGTGAC

20T ACCTATAGGGCTCGCGCCGACBTCVGTBCGBCGHTATACDACTACAGGAGCGAGTCGTAA

Fig. 19A

ENV BOTTOM OLIGOS

1B CCCCVCGTTTGTTACGACTCGCTCCTGTAGT

2B CTGGCGCGAGCCCTATAGGTGTKATTTCRATVAGCTTATAGTCACCGAGTTCGAGCCTGT

3B GACATCGTTATGTTCGTTTCGTTHGAGTTWATCCAGTCRATTTCAGCAATCAATGATGTA

4B ATCACCCTCGCGGGGTGGGAGGTAVACGTTCTTHCCVACCTTATGCCAGGTATTTATAAT

5B CGTAGTTCTTTTTTTGCTGTTCCTTDCGGTTCTGCTGCTTCCAACGCCCTCCTTTCTGAT

6B AAGAACCAGTTCATCTTACAGTAVAGAAACTCVCCYCTGCAGTTAGTCCACATGAATGTC

7B AGGTGCAGTCAAGTTTATCTTDCGBGTGTCGTTBGTVCCDGTGTAGCGCGGGTGCTTCAC

8B GTATCGCTTTCTTCCACGAACCCTCGAACCAACACCAHGCTTGCTTAGGACGTTCGTTAA

9B AGGCCACTCATTATCGTTACVGGVAGBACVGTTTTGTTVCCAGGGCGACGGCAACGCATA

10B GAGCGATATAATCGTACGGTTBGACTTHCCGTGCCAGTAKATATAAGTACGGTTCTCAGC

11B ACCATGTGCTCGTCTGGGTTTCCATCATVCGVGTACARCTGCTTACGACAACTTTACTGC

12B TAGTTACTGTCGTTGCAGCGYAADAGBGCGTABCCVGGDGGAGCGCAATAGCGGAAACGT

13B ATCGCAACTCTCCTGTATTACBGAHGTATTGCAGTGGTGCATATAGCACTTGGATTCTGA

14B ATATAAGGTCACGACTGTACCAVGTCTCGTTGTACTCCTTTTTCTTATCGCGGTTAAGGC

15B CCAATCATCGGTTCTTGCTCRAGHCCVGCGCABGAGTTGTTCTTAATACATGGGTTTGAC

16B AACTGAAGGAGTAGCGGTAGTVGTRATVGCVGTVGTVGTVGTACCAGCATTACCAGTGAG

17B TCTTGTTGCACCTCATAGCAATGCACAAVGGHGTYAACTTAACGCATGGTTTTATACTAG

18B TCTTCAATAGCCTGCTCAGTHACVGTGTTGTCCCABGCGTCGAATGCTTCAGTTACGTTT

19B GTCGTTGTCAGGGAGACACTGVGTVGTVCCCCABGTATCDCGGTTTTTTGTAGCGCAAAA

20B TCCAAGCAGGAACTCCGTAAAAVACVGTVACGTACTGBACACAACAAATCTCTAAAACAC

Fig. 19B

CORRECTED VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
C12N 15/10, 15/11, 5/10, 7/04, A61K 39/21, C07B 61/00, G06F 19/00, B01J 19/00 // C12N 15/12, 15/55, 15/49

A3

(11) International Publication Number: WO 00/18906

(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22588

(22) International Filing Date: 28 September 1999 (28.09.99)

(30) Priority Data:

60/102,362	29 September 1998 (29.09.98)	US
60/117,729	29 January 1999 (29.01.99)	US
60/118,813	5 February 1999 (05.02.99)	US
60/141,049	24 June 1999 (24.06.99)	US

(71) Applicant *(for all designated States except US)*: MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: PATTEN, Phillip, A. [US/US]; Apartment 506, 2680 Fayette Drive, Mountain View, CA 94040 (US). LIU, Lu [US/US]; 519 Skiff Circle, Redwood City, CA 94056 (US). STEMMER, Willem, P., C. [NL/US]; 108 Kathy Court, Los Gatos, CA 95030 (US).

(74) Agents: QUINE, Jonathan, Alan; Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(88) Date of publication of the international search report:
26 October 2000 (26.10.00)

(54) Title: SHUFFLING OF CODON ALTERED GENES

(57) Abstract

Methods of recombining codon–altered libraries of nucleic acids are provided. The nucleic acids can include conservative or non–conservative modifications of coding sequences, in addition to codon alterations, as compared with wild–type sequences. In addition to making new proteins, methods of generation vectors with reduced rates of reversion to wild–type and attenuated viruses are also provided.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

NSDOCID: <WO___0018906A3_I_>

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 99/22588

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12N15/10 C12N15/11 C12N5/10 C12N7/04 A61K39/21
C07B61/00 G06F19/00 B01J19/00 //C12N15/12,C12N15/55,
C12N15/49

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12N C07K A61K C07B G06F B01J

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 35966 A (MAXYGEN INC ;MINSHULL JEREMY (US); STEMMER WILLEM P C (US)) 2 October 1997 (1997-10-02) abstract page 25, line 1 - line 29 ---	1-49
X	WO 98 13485 A (ES HELMUTH H G VAN ;MAXYGEN INC (US); STEMMER WILLEM P C (US)) 2 April 1998 (1998-04-02) abstract page 9, line 28 -page 10, line 13 page 46, line 31 -page 47, line 34 ---	1-49
A	-/—	50,51

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 July 2000	28.07.00

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Galli, I

Form PCT/ISA/210 (second sheet), (July 1992)

BNSDOCID: <WO___0018906A3_I_>

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Idem as subject-matter 1, but wherein the target codon-altered nucleic acid is recombined with a portion of a viral genome to produce a viral vector that requires trans-complementation for replication and has a reduced rate of reversion to a replicative form.

8. Claims: (50,51) - completely

An integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acids strings and a user interface allowing a user to selectively view one or more sequence strings in the database.

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inte. .onal Application No
PCT/US 99/22588

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9735966 A	02-10-1997	US 5837458 A	17-11-1998
		AU 2337797 A	17-10-1997
		AU 2542697 A	17-10-1997
		CA 2247930 A	02-10-1997
		EP 0932670 A	04-08-1999
		EP 0906418 A	07-04-1999
		WO 9735957 A	02-10-1997
		AU 713952 B	16-12-1999
		AU 1087397 A	19-06-1997
		CA 2239099 A	05-06-1997
		EP 0876509 A	11-11-1998
		EP 0911396 A	28-04-1999
		JP 2000500981 T	02-02-2000
		WO 9720078 A	05-06-1997
WO 9813485 A	02-04-1998	AU 4503797 A	17-04-1998
		AU 4597197 A	17-04-1998
		EP 0964922 A	22-12-1999
		EP 0963434 A	15-12-1999
		WO 9813487 A	02-04-1998
WO 9206176 A	16-04-1992	AU 655788 B	12-01-1995
		AU 8922391 A	28-04-1992
		CA 2092803 A	29-03-1992
		EP 0551438 A	21-07-1993
		JP 6504666 T	02-06-1994
		NZ 239987 A	26-08-1994
		US 5770434 A	23-06-1998

BACKFILE DOCUMENT INDEX SHEET

A DOCPHOENIX

APPL PARTS

_____ IMIS ____
Internal Misc. Paper

_____ LET. ____
Misc. Incoming Letter

_____ 371P ____
PCT Papers in a 371Application

_____ A... ____
Amendment Including Elections

_____ ABST ____
Abstract

_____ ADS ____
Application Data Sheet

_____ AF/D ____
Affidavit or Exhibit Received

_____ APPENDIX ____
Appendix

_____ ARTIFACT ____
Artifact

_____ BIB ____
Bib Data Sheet

_____ CLM ____
Claim

_____ COMPUTER ____
Computer Program Listing

_____ CRFL ____
All CRF Papers for Backfile

_____ DIST ____
Terminal Disclaimer Filed

_____ DRW ____
Drawings

_____ FOR ____
Foreign Reference

_____ FRPR ____
Foreign Priority Papers

_____ IDS ____
IDS Including 1449

_____ NPL ____
Non-Patent Literature

_____ OATH ____
Oath or Declaration

_____ PET. ____
Petition

_____ RETMAIL ____
Mail Returned by USPS

_____ SEQLIST ____
Sequence Listing

_____ SPEC ____
Specification

_____ SPEC NO ____
Specification Not in English

_____ TRNA ____
Transmittal New Application

OUTGOING

_____ CTMS ____
Misc. Office Action

_____ 1449 ____
Signed 1449

_____ 892 ____
892

_____ ABN ____
Abandonment

_____ APDEC ____
Board of Appeals Decision

_____ APEA ____
Examiner Answer

_____ CTAV ____
Count Advisory Action

_____ CTEQ ____
Count Ex parte Quayle

_____ CTFR ____
Count Final Rejection

_____ CTNF ____
Count Non-Final

_____ CTRS ____
Count Restriction

_____ EXIN ____
Examiner Interview

_____ M903 ____
DO/EO Acceptance

_____ M905 ____
DO/EO Missing Requirement

_____ NFDR ____
Formal Drawing Required

_____ NOA ____
Notice of Allowance

_____ PETDEC ____
Petition Decision

INCOMING

_____ AP.B ____
Appeal Brief

_____ C.AD ____
Change of Address

_____ N/AP ____
Notice of Appeal

_____ PA.. ____
Change in Power of Attorney

_____ REM ____
Applicant Remarks in Amendment

_____ XT/ ____
Extension of Time filed separate

Internal

_____ SRNT ____
Examiner Search Notes

_____ CLMPTO ____
PTO Prepared Complete Claim Set

_____ ECBOX ____
Evidence Copy Box Identification

_____ WCLM ____
Claim Worksheet

_____ WFEE ____
Fee Worksheet

File Wrapper

_____ FWCLM ____
File Wrapper Claim

_____ IIFW ____
File Wrapper Issue Information

_____ SRFW ____
File Wrapper Search Info

REV 7/02/03

Evolution of a cytokine using DNA family shuffling

Chia-Chun J. Chang[#], Teddy T. Chen[#], Brett W. Cox, Glenn N. Dawes, Willem P.C. Stemmer, Juha Punnonen, and Phillip A. Patten[*]

Maxygen, Inc., 3410 Central Expressway, Santa Clara, CA 95051. [#]These authors contributed equally to this work. [*]Corresponding author (e-mail: phil_patten@maxygen.com).

Received 22 February 1999; accepted 21 May1999

DNA shuffling of a family of over 20 human interferon-α (Hu-IFN-α) genes was used to derive variants with increased antiviral and antiproliferation activities in murine cells. A clone with 135,000-fold improved specific activity over Hu-IFN-α2a was obtained in the first cycle of shuffling. After a second cycle of selective shuffling, the most active clone was improved 285,000-fold relative to Hu-IFN-α2a and 185-fold relative to Hu-IFN-α1. Remarkably, the three most active clones were more active than the native murine IFN-αs. These chimeras are derived from up to five parental genes but contained no random point mutations. These results demonstrate that diverse cytokine gene families can be used as starting material to rapidly evolve cytokines that are more active, or have superior selectivity profiles, than native cytokine genes.

Keywords: DNA shuffling, cytokines, interferon, molecular breeding, pharmaceutical proteins

Alpha interferons (IFN-αs) are members of the diverse helical-bundle superfamily of cytokine genes[1]. Although these proteins possess therapeutic value in the treatment of a number of diseases, they have not been optimized for use as pharmaceuticals. For example, dose-limiting toxicity, receptor cross-reactivity, and short serum half-lives significantly reduce the clinical utility of many of these cytokines[2–5].

The human IFN-αs (Hu-IFN-αs) are encoded by a family of over 20 tandemly duplicated nonallelic genes that share 85–98% sequence identity at the amino acid level[6]. These proteins have potent antiviral and antiproliferative activities that have clinical utility as anticancer and antiviral therapeutics. Although the utility of chimeric IFNs derived from this gene family has been recognized[7], only a small fraction of the 10^{26} possible chimeras have been explored either in natural human evolution or by the methods of modern molecular biology; and only one natural IFN-α subtype, Hu-IFN-α2, has been used in clinical studies[7]. The most active engineered IFN-α, IFN alfacon-1, is a consensus of 13 wild-type Hu-IFN-α genes that is currently used in hepatitis C therapy[8].

DNA family shuffling is a method for permutation of natural genetic diversity. It is a powerful tool for rapidly evolving genes, operons, and whole viruses for desired properties[9–11]. In this report, we describe the rapid evolution of the Hu-IFN-α genes for activity in mouse cells using DNA family shuffling. The native Hu-IFN-α genes are 53–65% identical to murine (Mu) IFN-αs and exhibit very weak activity on murine cells[7]. Similarly, the extracellular domains of the IFN-α receptors share only 49% sequence identity[12]. Despite these sequence differences, we have obtained shuffled IFN-αs that are more potent in mouse cells than the native Mu-IFN-αs.

Results and discussion

A primary round of DNA family shuffling was performed to generate a library of chimeric Hu-IFN-αs. All of the Hu-IFN-α genes, including pseudogenes, were shuffled in order to capture the diversity of the entire family. Chimeric IFN-αs were expressed, purified, and screened for L929 antiviral activity as pools of 12 or 96. The active pools were deconvoluted into sequentially smaller pools to identify single active clones. Because a pooling strategy was used, only 68 murine antiviral assays were used to screen a library of 1,672 clones. The most active chimeric IFN-α from round 1 (IFN-CH1.1) was derived from six parental Hu-IFN-α gene segments (Fig. 1), and is 87-fold more active than Hu-IFN-α1, the most active native human interferon, in murine cells (Table 1).

DNA family shuffling mimics and extends classical breeding methods by recombining more than two parental genes, or genes from different species, in a single DNA shuffling reaction[9–11]. The power of breeding is that large improvements in phenotype can be achieved by recursively screening only a small subset of all theoretically possible progeny[13,14]. It is therefore important to determine the most efficient and selective breeding strategies.

In a second round of shuffling, we compared pooled and pair-wise matings of the four IFN-α genes from round 1 to make five new libraries of chimeras. Four libraries were made by pair-wise matings of the genes and one library by pooled mating of all four genes. A high-throughput assay was used to screen 1,056 individual clones from this panel of five libraries, and the top 60 candidates were rescreened quantitatively for antiviral activity in L929 cells. The genes from the 11 most active shuffled IFN-αs were expressed in Chinese hamster ovary (CHO) cells, and purified IFN-α protein was assayed. The most active IFN-α from cycle 2 is improved 185-fold relative to Hu-IFN-α1 and 285,000-fold relative to Hu-IFN-α2a (Figs 2 and 3). Remarkably, the activities of the three most active IFN-αs exceed the activity of the most active native mouse IFN-α, Mu-IFN-α4 (Table 1, Fig. 3). The most active clones from round 2 came from the pair-wise matings of highly active clones (IFN-CH1.1 × IFN-CH1.3), with none of the most active clones in round 2 coming from the pooled mating (Table 1). The superior performance of pair-wise matings relative to pooled matings may reflect sparse sampling of a population with a significantly lower average level of biological activity in clones derived from the pooled mating, due to breaking up favorable amino acid combinations such as Lys121 and Arg125, as discussed below.

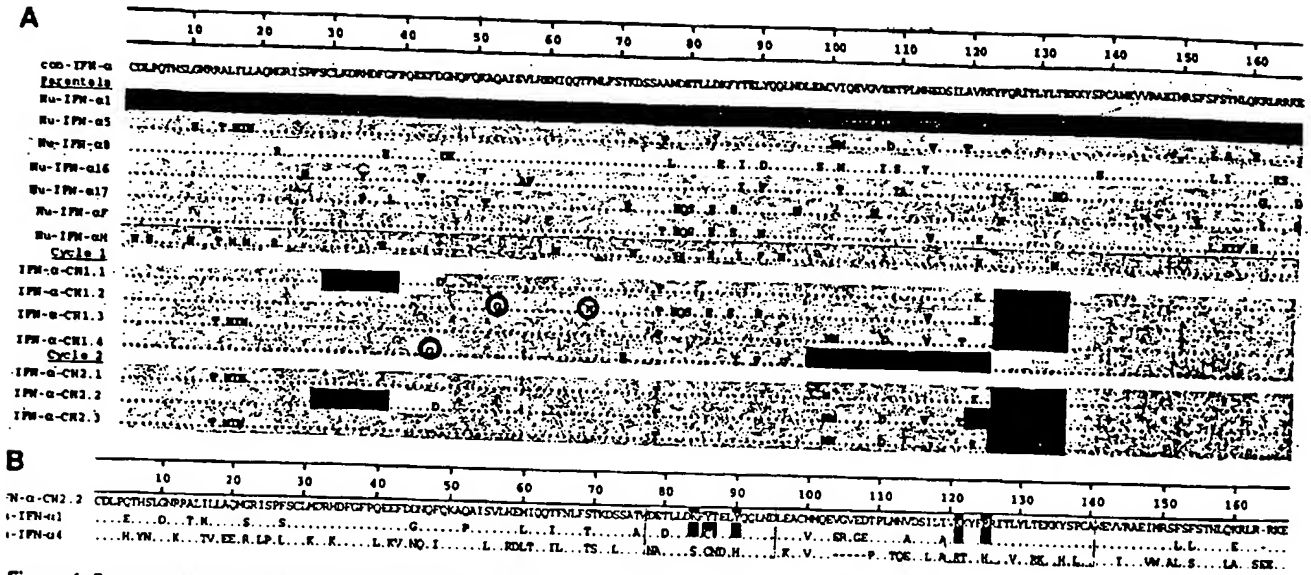


Figure 1. Sequences and genealogies of shuffled interferons. (A) The amino acid sequences of seven evolved IFN-αs and the eight native Hu-IFN-αs from which they are derived. The most parsimonious genealogies of the shuffled IFN-αs are shown schematically. Recombination junctions are shown at the midpoint between two amino acids derived from different parental genes. The gene segments are colored according to which parental gene they are derived from (red, Hu-IFN-α1; green, Hu-IFN-α5; yellow, Hu-IFN-α8; purple, Hu-IFN-α16; orange, Hu-IFN-α17; blue, Hu-IFN-αF; gray, Hu-IFN-αH). Amino acids that arose by point mutation during DNA shuffling are circled. (B) The sequence of one of the cycle 2 chimeras, IFN-CH2.2, is aligned with the most potent human and mouse IFN-αs, Hu-IFN-α1 and Mu-IFN-α4. The IFN-α residues that putatively contact the IFN-α receptor[27,28] are boxed. Residues in Hu-IFN-α1 that have been shown by site-directed mutagenesis to contribute to activity on mouse cells[7,25-28] are shaded.

Our libraries of chimeric interferons contained few inactive or weakly active clones. In contrast, random mutagenesis leads frequently to gene inactivation[15,16]. For example, 75% of random point mutants of residues 120–136 of Hu-IFN-α4 are inactive[17]. To assess the knockout rate in our primary libraries, we assayed four randomly chosen intact IFN-α chimeric clones in a human cell proliferation assay (using Daudi cells). All four shuffled IFN-αs are as active in human Daudi cells as Hu-IFN-α2a, despite having 10–21 amino acid changes relative to the closest native Hu-IFN-α[6] (Fig. 1; Table 1). Two thirds of the clones from the second round of DNA family shuffling are more active in mouse cells than Hu-IFN-α1, the most active native Hu-IFN-α1 (C.-C.J.C. and T.T.C., unpublished data). The diversity in the libraries in this study was overwhelmingly generated by recombination of preexisting natural sequence diversity in the gene family, with random point mutation accounting for only two sequence changes in the four round 1 chimeras (Fig. 1). These random mutations were removed in the second round of breeding by recombination with native gene segments; thus, there were no random point mutations in the three most active round 2 chimeras (Fig. 1). These results provide an additional indication of the high quality of libraries made by DNA family shuffling.

DNA family shuffling permutes blocks of sequence containing conservative amino acid substitutions that have been selected for function during millions of years of natural selection[9-11,15]. Consequently, the sequence space defined by recombination of natural diversity is highly preselected for function and represents only a

Table 1. Activities of parental and evolved IFN-αs in murine cells[a].

DNA shuffling cycle	IFN-α gene	Genealogy	L129 antiviral activity (U/mg × 10^{-6})	Fold improvement in activity versus Hu-IFN-α2a	Human cell anti-proliferation activity (U/mg × 10^{-6})
0	IFN-αCon-1	Synthetic	1.0	N/A	46
0	Hu-IFN-α2a	Native	0.00194	N/A	27
0	Hu-IFN-α1	Native	3.0	N/A	70
0	Mu-IFN-α1	Native	140	N/A	<0.001
0	Mu-IFN-α6	Native	116	N/A	<0.001
0	Mu-IFN-α4	Native	160	N/A	<0.001
1	IFN-CH1.4	IFN-α1, H, F, 17, 16	0.15	77	ND
1	IFN-CH1.3	IFN-α1, 5	16	8,247	20
1	IFN-CH1.2	IFN-α1, F	42	21,649	ND
1	IFN-CH1.1	IFN 1, 5, 8, 14, F	262	135,000	56
2	IFN-CH2.3	IFN-CH1.1 × IFN-CH1.3	268	138,000	62
2	IFN-CH2.2	IFN-CH1.1 × IFN-CH1.3	400	206,000	86
2	IFN-CH2.1	IFN-CH1.1 × IFN-CH1.3	554	285,000	62

[a]Native and evolved IFN-α genes (Fig. 1) were expressed in CHO cells[31], and purified protein was assayed for antiviral activity on L929 mouse cells. One unit of activity is defined as the amount of IFN-α required to give half-maximal protection against 100 $TCID_{50}$ of encephalomyocarditis virus. One unit per ml of activity on Daudi cells is defined as the concentration of protein required to give half-maximal inhibition of ^{3}H-thymidine incorporation. The genealogies of the cycle 1 and cycle 2 shuffled IFN-αs are indicated. The standard errors were between 10% and 32%. The fold improvement relative to Hu-IFN-α2a is indicated. N/A, not applicable.

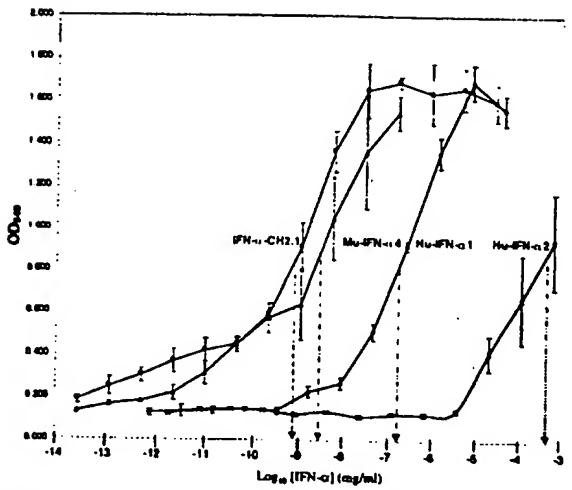


Figure 2. Antiviral activities of native IFN-αs and an evolved IFN-α. The results from the L929 antiviral assay of Hu-IFN-α2a, Hu-IFN-α1, Mu-IFN-α4, and IFN-CH2.1 are shown. The dashed lines indicate the IFN-α dose corresponding to half-maximal protection (1 U/ml). The assays were done in triplicate and the standard errors (percentage of the estimated units; Table 1) are: Mu-IFN-α4, 24%; Hu-IFN-α1, 6%; Hu-IFN-α2a, 17%; and IFN-CH2.1, 15%.

fraction of the sequence space accessible by random mutation. Thus, although the Hu-IFN-α gene family is variable at 76 sites[6], there is a very limited range of amino acid changes at these sites. In the natural molecule, there are two, three, or four amino acid alternatives at 57, 15, and 4 sites, respectively[6], so the number of possible recombinants is $2^{57} \times 3^{15} \times 4^4 = 5 \times 10^{26}$. As the Hu-IFN-αs differ by an average of 17 out of 166 amino acid residues[6], there are 166!/(166 − 17)!(17!) = 7×10^{22} ways to distribute 17 mutations over 166 residues, and for any given distribution $20^{17} = 1.3 \times 10^{22}$ possible random mutants. A simple multiplication reveals 9×10^{44} possible 17-step random mutants of a 166-residue protein such as Hu-IFN-α. Thus, there are ~10^{45} 17-step random mutants of Hu-IFN-α, whereas there are only 10^{26} permutations of the natural Hu-IFN-α sequence diversity. In effect, shuffled IFN-αs sample only 10^{-19} of the random point-mutant spectrum. In contrast to family shuffled libraries, an infinitesimal fraction of 17-step random point mutants of Hu-IFN-αs are expected to be active[15–17]; therefore, these libraries of shuffled chimeras are highly enriched for functional clones relative to libraries made by random point mutagenesis. This result illustrates the ability of DNA family shuffling to generate progeny that differ from the parent molecules at many residues, while still retaining potent biological activity.

As a consequence of the high average activity of members of the family shuffled libraries, direct screening for biological activity is possible. This is a significant advantage over other strategies such as phage panning, in that a small number of complex biological assays can be used to directly obtain clones with the desired biological activity. Thus, evolution of commercially important genes and proteins may be practical, even when very complex, time-consuming, or expensive assays are required.

Immunogenicity is clinically significant for many recombinant pharmaceutical proteins[18–20]. We expect that because of the functionally conservative nature of family shuffling, breeding closely related gene homologs, rather than performing random or site-directed mutagenesis, is more likely to generate immunologically conservative chimeras. Undesired T- and B-cell epitopes can be removed from shuffled clones by back-crossing evolved IFN-αs with wild-type IFN-αs and screening for genes that retain high activity but lose immunogenic epitopes.

Classical inbreeding to enhance a particular phenotype can result in loss of characteristics in the parentals that are not under selective pressure[21]. In this study, we selectively bred for activity in murine cells, with no pressure for retention of activity on the Hu-IFN-α receptor, which is only 49% identical in amino acid sequence[12]. Surprisingly, all of these clones retained antiproliferative activity in human cells that is within twofold of the activity of Hu-IFN-α2a (Table 1), whereas none of the Mu-IFN-αs has detectable activity in human cells (<10^{-5} of the activity of Hu-IFN-α). This illustrates how DNA family shuffling of natural gene families, by using recombination of functionally conservative natural sequence diversity within a gene family rather than random point mutation, can allow one to evolve cytokines that retain activity on one receptor while gaining activity on a homologous receptor. The ability to evolve pluripotent cytokines may be useful in the development of novel protein therapeutics, such as for proteins active in multiple plants, farm animals, or pathogens.

Previous engineering of cytokines has relied principally on site-directed mutagenesis guided by structural models[22] and on cassette mutagenesis or random mutagenesis[23,24]. Improving genes by classical structure/function analysis generally relies on measuring the effect of single mutations or cassettes of mutations in one context, and then multistep mutants are built up based on the presumed additive nature of these mutations[22–24]. Consequently, combinations of mutations that have nonadditive effects are difficult to discover[25]. Several studies have identified residues in chimeric and point-mutated Hu-IFN-αs that confer activity in murine cells. Replacing residues 61–92 of Hu-IFN-α8 with those from Hu-IFN-α1 significantly increases the activity in murine cells, and point mutagenesis implicates residues 84, 86, 87, and 90 as contributing to this effect[7]. Analysis of a series of 20 chimeras between Hu-IFN-α1 and Hu-IFN-α2a reveals that sequences in the C-terminal 49 residues are responsible for its unusually high activity in murine cells[26]. Further analysis by site-directed mutagenesis reveals that transfer of residues Lys121 or Arg125 to Hu-IFN-α2 increases activity on murine cells, and that together they increase activity by 400-fold[26]. Based on these functional data and on homology modeling, the residues in these two regions (78–95 and 121–132) have been proposed to interact with the Mu-IFN-α receptor[27,28].

Interestingly, Lys121 and Arg125 occur either separately or together in all of our cycle 1 chimeras, and both residues occur together in all five of the most active chimeras from cycle 2 (Fig. 1). While the three most active chimeras are identical to Hu-IFN-α1 at five of the six

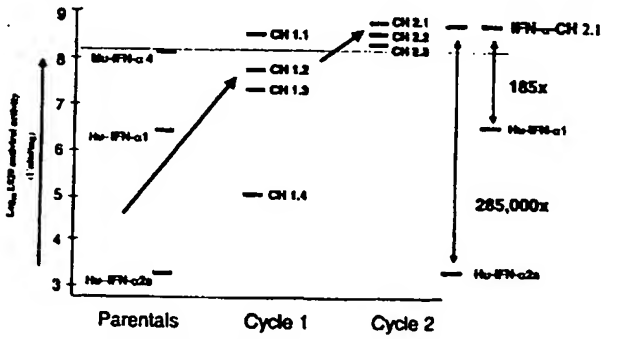


Figure 3. Summary of antiviral activities of native and evolved IFN-αs on murine L929 cells. The antiviral activities of purified CHO protein for native Mu-IFN-αs, native Hu-IFN-αs and evolved IFN-αs on murine L929 cells are shown. One unit of activity corresponds to half-maximal protection from a lethal ECMV viral challenge. The arrows on the right indicate the fold improvement of IFN—CH2.3 relative to Hu-IFN-α1 and Hu-IFN-α2a. The activities of the proteins were measured in four independent experiments, and the rank orders of the clones was the same in all four assays, with the exception of assay 3 in which Mu-IFN-α4 exceeded the activity of IFN—CH1.1, but not the round 2 evolved IFN-αs.

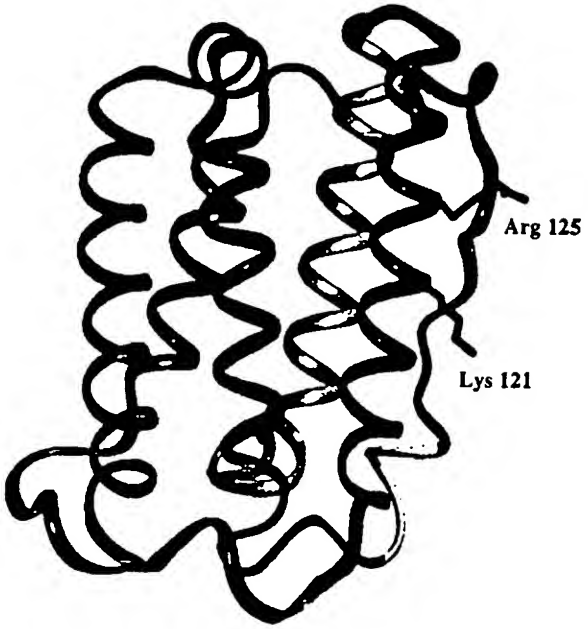


Figure 4. Structural modeling of IFN-CH2.2. A model of the C_α backbone of IFN-CH2.2, based on the NMR structure of Hu-IFN-α2a (ref. 30), is shown. The protein backbone is colored to indicate the native Hu-IFN-α segment from which it is derived (red, residues 29–39, 121–140 Hu-IFN-α1; green, residues 46–120 Hu-IFN-α5; yellow, residues 40–45 Hu-IFN-α8; blue, residues 1–28 Hu-IFN-αF; gray, residues 141–166 Hu-IFN-αH). The side chains of putative murine IFN-α receptor contacting residues Lys121 and Arg125 are shown.

residues that have previously been shown to contribute to its activity in mouse cells[7,26], they contain 22–28 additional sequence changes relative to Hu-IFN-α1 (Fig. 1). This large number of differences from the parental genes is typical of chimeras constructed by DNA family shuffling because blocks of sequence are shuffled; thus, progeny sequences generally have many amino acid differences from the closest parental molecules. As a result, complex improvements need not be built up in multiple rounds of low-level mutation or by using powerful selection methods on large libraries. These clones are improved by up to 285,000-fold relative to Hu-IFN-α2a, an additional 500-fold increase in activity relative to the Lys121, Arg125 double mutant (Hu-IFN-α2a plus Lys121 and Arg125 from Hu-IFN-α1; ref. 26). The three most active chimeras in this report are more active in murine cells than any chimeras or point mutants reported in any previous studies[7,26], and are the first examples of Hu-IFN-α variants that are more active than the native Mu-IFN-αs. This study illustrates the utility and novel aspects of DNA family shuffling for recruiting, from gene families, segments of genes that confer or enhance a novel biological activity, and for sequentially optimizing them, without a priori guidance from structural or functional information. Lys121 and Arg125 occur on the solvent-exposed face of one helix (Fig. 4). The prominence of these residues in our study, taken together with previous work[26–28], suggests that this helix may contact the murine IFN-α receptor.

In summary, DNA family shuffling of IFN-α genes from one species and a modest number of cell-based assays allowed us to rapidly obtain recombinants with potent IFN-α activity on a distantly related species. This result has broad implications for cytokine engineering, suggesting that diverse mammalian homologs of human cytokines can be used as breeding stock from which to evolve cytokines that are more active or have superior selectivity profiles than native cytokine genes. For example, it may be possible to evolve Hu-IFN-αs with reduced side effects[2–5], improved antitumor activity in humans[29], or IL-2 variants with reduced toxicity[2].

Experimental protocol

DNA cloning, sequencing, and shuffling. The Hu-IFN-α gene family was PCR amplified from human genomic DNA using 12 sets of degenerate primers (sequences available upon request). DNase I was used to fragment 300 μg of PCR product; 25–60 bp fragments were gel purified and shuffled as described[11]. Two additional libraries of shuffled Hu-IFN-α genes were made from eight cloned Hu-IFN genes (Hu-IFN-αs 1, 4, 5, 6, 14, 16, 17, and F). Assembled insert was introduced into the phagemid display vector pDEI-932. Hu-IFN alfacon-1 was constructed from synthetic oligonucleotides. Hu-IFN-αs 1, 2a, 4, 5, 6, 14, 16, 17, and F; and Mu-IFN-αs 1, 4, and 6 were cloned from genomic DNA and sequenced on an ABI DNA sequencer (Perkin Elmer, Foster City, CA). The extracellular domains of the human and mouse IFN-α receptors were aligned by the Clustal method (DNA Star, Madison, WI; SWISS-PROT database accession nos. P33896, P17181, and P48551; GENBANK accession no. AF013274).

Phagemid display of IFN. Shuffled Hu-IFN-α genes were expressed by the gene III phagemid display vector pDEI-932, comprising an STII leader fused to the N terminus of Hu-IFN-α and the E-tag (Amersham Pharmacia, Piscataway, NJ) with a 6-His tag at the C terminus. Following the C-terminal tag is a suppressible amber codon, followed by M13 gene III (fused at residue 247 of gene III). The IFN-α gene III insert is under the control of the pBAD promoter, and the backbone plasmid is an Amp^R derivative of pBR322 containing an M13 origin of replication. Large-scale (250 ml) phagemid preparations were done by standard methods[31] in the presence of 0.002% arabinose to induce expression of the IFN-α gene III fusion. Phagemids were PEG precipitated, CsCl banded, and dialyzed into PBS before assaying.

High-throughput phagemid preparations. For the purposes of HTP screening, *Escherichia coli* harboring phagemids were picked with a Q-BOT robotic colony picker (Genetix Pharmaceuticals, Tarrytown, NY) into 96-well plates containing 100 μl of 2xYT per well. Confluent cultures were grown overnight at 37°C. The overnight cultures were diluted 20-fold into fresh 2xYT, Amp/0.002% arabinose wt/vol/10^{10} p.f.u./ml M13 VCS helper phage and grown for 4 h with vigorous shaking. The cells were pelleted, and phage supernatants were transferred to 96-well dialysis plates containing a 100 kDa cutoff membrane before assaying. Samples were dialyzed against PBS and then filter sterilized through 96-well 0.45 μm membranes. Sterile phagemid samples were used directly in cellular assays.

Antiviral assays. Antiviral activities were determined by the cytopathic effect (CPE) reduction assay on mouse L929 cells challenged with encephalomyocarditis virus (EMCV). Briefly, target cells were grown to confluence, trypsinized, and distributed into 96-well flat-bottom microtiter plates (10^4 cells/well) in RPMI medium supplemented with 10% fetal calf serum and penicillin/streptomycin antibiotics. IFN-α samples were titrated in triplicate in five-fold dilutions. After incubation for 16 h, the medium was removed, replaced with medium containing EMCV (100 $TCID_{50}$ per well), and the plates were incubated for two days until CPE occurred. Medium was removed, the cells were washed twice with PBS, and neutral red (1:100 dilution) was added, followed by incubation for 2 h. During the last 20 min, cells were fixed with 0.5% glutaraldehyde. The unstained dye solution was removed, the plates were washed twice with PBS, and the color was extracted with 50% methanol, 1% acetic acid. The extracted dye solution in the well was quantitated colorimetrically at 540 nm with a spectrophotometer. Results of the CPE reduction assay derived as above were plotted to produce sigmoidal dose–reponse curves by plotting the logarithm of the IFN-α concentration versus the cell viability. One unit per milliliter is defined as the interpolated IFN-α concentration giving 50% protection (on a scale of 0 to 100% determined by controls with no IFN-α and with or without virus).

Deconvolution of libraries. In cycle 1, eight pools of 12 were assayed, and one had measurable antiviral activity. Sixteen pools of 96 were assayed, the most active pool of 96 was broken into eight pools of 12, and these pools were assayed separately. Three pools of 12 had measurable activity, and 36 individual phagemids were prepared, purified, and assayed from these pools. One chimera (Hu-IFN-CH1.4) was obtained by randomly screening individual clones in the library for L929 antiviral activity. Three IFN-α phagemids with antiviral activity were obtained (one from each pool). The IFN-α chimeras from these phagemids were cloned into the CHO expression vector pDEI-1011, transfected, and purified as described.

Construction of round 2 libraries. Five cycle 2 libraries were constructed by shuffling equimolar qualities of plasmid DNA in the following combinations: IFN-CH1.1 × IFN-CH1.2; IFN-CH 1.1 × IFN-CH1.3; IFN-CH1.1 × IFN-CH1.4; IFN-CH 1.2 × IFN-CH 1.3; IFN-CH 1.1 × IFN-CH 1.2 × IFN-CH 1.3 × IFN-CH 1.4. Shuffled libraries were made in pDEI-932 from 25–50 bp fragment assemblies, as described[11].

High-throughput proliferation assays. The L929 antiproliferative assay was performed according to standard [^{3}H]thymidine incorporation methods. Briefly, IFN-α samples were titrated in triplicate in fivefold dilution steps down the plate. For high-throughput screening in the second round of shuffling, four single 10-fold dilutions were assayed in the primary screen, and subsequent rescreens were done in triplicate. L929 cells (1,000/well) were incubated for 72 h at 37°C, 5% CO_2. During the last 16 h of incubation, 1 μCi/well of [^{3}H]thymidine was added. The plates were then harvested on a Harvester-96 (Tomtec Inc., Orange, CT), and thymidine incorporation was counted on a beta counter (Microbeta, Wallac, Gaithersburg, MD).

CHO expression and purification of chimeric IFN-αs. IFN-α genes were cloned into a standard CHO expression vector (pDEI-1011) in which the E-tag/6-His tag (Pharmacia) is fused to the C terminus of the IFN-αs. Expression is driven by the SR-α promoter, and stable transfectants were selected at 1 mg/ml G418. The four most active clones from the first round and the 15 most active clones from the second round were inserted into a pDEI-1011 and introduced into CHO cells by transfection[31]; the proteins were affinity purified from the supernatant on anti-E tag Sepharose (Pharmacia).

Daudi proliferation assays. Eight chimeric phage-displayed IFN-αs were sequenced from randomly picked clones. Four of the eight sequences encoded in-frame IFN-α genes. These four chimeras and Hu-IFN-α2a were expressed, purified, and assayed for antiproliferation activity on human Daudi cells, as described[32]. One unit per milliliter is defined as the concentration giving half-maximal inhibition of proliferation. Two thirds of the clones in the cycle 2 library are more potent than Hu-IFN-CH1.4 in the HTP 1.929 antiproliferation assay.

Acknowledgments

We thank Peter Schultz, Mark Gallop, Eleanor Fish, Russell Howard, Jennifer Jones, Steve Bass, and Volker Heinrichs for critical reading of the manuscript. We thank Malcolm McGregor for assistance with the molecular graphics and Andreas Crameri for sharing unpublished DNA shuffling protocols.

1. Sprang, S.R. & Bazan, J.F. Cytokine structural taxonomy and mechanisms of receptor engagement. *Curr. Opin. Struct. Biol.* **3**, 815–827 (1993).
2. Dusheiko, G. Side effects of alpha interferon in chronic hepatitis C. *Hepatology* **26**, 112S–121S (1997).
3. Vial, T. & Descotes, J. Clinical toxicity of the interferons. *Drug Experience* **10**, 115–150 (1994).
4. Funke, I. *et al.* Capillary leak syndrome associated with elevated IL-2 serum levels after allogeneic bone marrow transplantation. *Ann. Hematol.* **68**, 49–52 (1994).
5. Schomburg, A., Kirchner, H. & Atzpodien, J. Renal, metabolic and hemodynamic side-effects of interleukin-2 and /or interferon alpha: evidence of a risk /benefit advantange of subcutaneous thereapy. *J. Cancer Res. Clin. Oncol.* **119**, 745–755 (1993).
6. Henco, K. *et al.* Structural relationship of human interferon alpha genes and pseudogenes. *J. Mol. Biol.* **185**, 227–260 (1985).
7. Horisberger, M.A. & Di Marco, S. Interferon-alpha hybrids. *Pharmacol. Ther.* **66**, 507–534 (1995).
8. Blatt, L.M., Davis, J.M., Klein, S.B. & Taylor, M.W. The biologic activity and molecular characterization of a novel synthetic interferon-alpha species, consensus interferon. *J. Interferon Cytokine Res.* **16**, 489–499 (1996).
9. Stemmer, W.P.C. Searching sequence space. *Biotechnology* **13**, 549–555 (1995).
10. Patten, P.A., Howard, R.H. & Stemmer, W.P.C. Applications of DNA shuffling to pharmaceuticals and vaccines. *Curr. Opin. Biotechnol.* **8**, 724–733 (1996).
11. Crameri, A., Raillard, S.A., Bermudez, E. & Stemmr, R.W.P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **15**, 288–291 (1998).
12. Uze, G., Lutfalla, G. & Mogensen, K.E. Alpha and beta interferons and their receptor and their friends and relations. *J. Interferon Cytokine Res.* **15**, 3–26 (1995).
13. Burbank, L. in *Luther Burbank, his methods and discoveries: their practical application* Vol. 1 (eds Whitson, J. & Williams, R.J.H.S.) 176–210 (Luther Burbank Press, New York; 1914).
14. Haldane, J.B.S. A mathmatical theory of natural and artificial selection *Cambridge Phil. Soc. Trans.* **23**, 19–41 (1924).
15. Muller, H.J. The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2–9 (1964).
16. Moore, J.C., Jin, H. & Arnold, F.H. Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336–347 (1997).
17. Tymms, M.J., McInnes, B., Alin, P., Linnane, A.W. & Cheetham, B.F. Structure-function studies of interferon-alpha based on random mutagenesis and expression *in vitro*. *Genet. Anal. Techn. Appl.* **7**, 53–63 (1990).
18. van der Meide, P.H. & Schellekens, H. Anti-cytokine autoantibodies: epiphenomenon or critical modulators of cytokine action. *Biotherapy* **10**, 39–48 (1997).
19. Konrad, M. Immunogenicity of proteins administered to humans for therapeutic purposes. *Trends Biotechnol.* **7**, 175–179 (1989).
20. Allegreta, M. *et al.* The development of anti-interleukin-2 antibodies in patients treated with recombinant human interleukin-2 (IL-2). *J. Clin. Immunol.* **6**, 481–490 (1986).
21. Lynch, M. & Wallace, B. in *Genetics and analysis of quantitative traits* (Sinauer Associates, Sunderland, MA; 1998)
22. Fuh, G. *et al.* Rational design of potent antagonists to the human growth hormone receptor. *Science* **256**, 1677–1680 (1992).
23. Lowman, H.B. & Wells, J.A. Affinity maturation of human growth hormone by monovalent phage display. *J. Mol. Biol.* **234**, 564–578 (1993).
24. Thomas, J.W. *et al.* Potent interleukin 3 receptor agonist with selectively enhanced hematopoietic activity relative to recombinant interleukin 3. *Proc. Natl. Acad. Sci. USA* **92**, 3779–3783 (1995).
25. Wells, J.A. Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509–8517 (1990).
26. Weber, H., Valaenzuela, D., Lujber, G., Gubler, M. & Weissmann, C. Single amino acid changes that render human IFN-alpha 2 biologically active on mouse cells. *EMBO. J.* **6**, 591–598 (1987).
27. Fish, E.N. Definition of receptor binding domains in interferon-alpha. *J. Interferon Res.* **12**, 257–266 (1992).
28. Uze, G. *et al.* Domains of interaction between alpha interferon and its receptor components. *J. Mol. Biol.* **243**, 245–257 (1994).
29. Gutterman, J.U. Cytokine therapeutics: lessons from interferon alpha. *Proc. Natl. Acad. Sci. USA.* **91**, 1198–1205 (1994).
30. Klaus, W., Gsell, B., Labhardt, A.M., Wipf, B. & Senn, H. The three-dimensional high resolution structure of human interferon alpha-2a determined by heteronuclear NMR spectroscopy in solution. *J. Mol. Biol.* **274**, 661–675 (1997).
31. Sambrook, J., Fritsch, E.F. & Maniatis, T. in *Molecular cloning: a laboratory manual* (Cold Spring Harbor Laboratory Press, New York; 1989).
32. Scarozza, A.M., Collins, T.J. & Evans, S.S. DNA synthesis in nuclei isolated from Daudi B cells: a model to study the antiproliferative mechanisms of interferon alpha. *J. Interferon Res.* **12**, 35–42 (1992).